HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

BACHELORARBEIT

Herr B.Sc
Mathias Langer

Analyse der molekularbiologischen Evolution der BRCT-Domäne im Energieprofil-orientierten Kontext

Mittweida, 2012

Fakultät Mathematik/ Naturwissenschaften/ Informatik

BACHELORARBEIT

Analyse der molekularbiologischen Evolution der BRCT-Domäne im Energieprofil-orientierten Kontext

Autor:
Herr B.Sc. Mathias Langer

Studiengang:
Biotechnologie/ Bioinformatik

Seminargruppe:
BI09w2-B

Erstprüfer:
Prof. Dr. Drik Labudde

Zweitprüfer:
B.Sc. Florian Heinke

Einreichung:
Mittweida, 17.08.2012

Verteidigung/Bewertung:
Mittweida, 2012

Bibliografische Beschreibung:

Langer, Mathias:
Analyse der molekularbiologischen Evolution der BRCT-Domäne im Energieprofil-orientierten Kontext . - 2012. – III und IV, 88, A-I S.
Mittweida, Hochschule Mittweida, Fakultät Fakultät Mathematik/ Naturwissenschaften/ Informatik, Bachelorarbeit, 2012

Referat:

Die vorliegende Arbeit liefert eine ausführliche Analyse der BRCT-Domäne. Dabei wurde zum einen eine Zusammenfassung erstellt, in welcher die wichtigsten bisher ermittelten bekannten Eigenschaften und Funktionen dieser Domäne zusammengetragen sind. Des Weiteren erfolgte eine Analyse der BRCT-Domäne auf Grundlage von Energieprofilen. Der Schwerpunkt der Untersuchung lag dabei auf der Rekonstruktion evolutionärer Verwandtschaftsbeziehungen. Somit wurde zum einen überprüft in wie weit sich Energieprofile auf phylogenetische Methoden anwenden lassen, wie deren Ergebnisse zu bewerten sind und in wie weit diese Ergebnisse mit bisherigen evolutionären Erkenntnissen der BRCT-Domäne korrelieren oder sich neue Erkenntnisse daraus ergeben.

Inhalt

Inhalt I

Abbildungsverzeichnis III

Tabellenverzeichnis V

1 Einleitung 1

1.1 Motivation 1

1.2 Zielsetzung 2

2 Grundlagen Proteine 3

2.1 Proteine Allgemein 3

2.2 Aufbau und Struktur von Proteinen 4

2.3 Proteinfaltung 7

3 Grundlagen phylogenetischer Methoden 9

3.1 Phylogenie Allgemein 9

3.2 Die UPGMA Methode 12

3.3 Die Neighbor-Joining Methode 14

3.4 Maximum Likelihood 19

4 Die BRCT-Domäne 24

4.1 Allgemeines 24

4.2 Aufbau und sequenzielles Verhalten 25

4.3 DNA-Bindende Funktion der BRCT-Domäne am Beispiel des Menschlichen Replikationsfaktor C p140 29

4.4 Protein bindende Funktion durch Dimerisierung zweier BRCT-Domäne am Beispiel der Bindung von XRCC1 mit der DNA Ligase III 32

4.5 Protein bindende Funktion durch Interaktion eines Proteins mit der BRCT-Domäne eines andere Proteins anhand des Beispielkomplexes XRCC4/DNA Ligase IV 37

4.6 Proteinbindende Funktion durch Bindung eines phosphoreszierten Peptides 40

4.7 Evolution der BRCT-Domäne 44

5 Theorie der Energieprofile 47

6 Konstruktion der zu Untersuchenden Stammbäume 52

6.1 Anwendung von Energieprofilen auf phylogenetischen Methoden 55

7 Untersuchung der Korrelation zwischen Stammbäumen auf sequenzieller, struktureller und energetischer Grundlage 58

8 Untersuchung des Verlaufes der Evolution der BRCT-Domänen auf energetischer und struktureller Ebene 64

8.1 Analyse der N-terminalen Double BRCT-Domänen 66

8.2 Analyse der C-terminalen Double BRCT-Domänen 70

8.3 Analyse der Single BRCT-Domänen 74

8.4 Rückschlüsse der energetischen Analyse auf den evolutionären Verlauf der BRCT-Domäne 83

9 Zusammenfassung 85

9.1 Ausblick 86

Literatur 87

Danksagung 92

Anlagen 93

Anlagen, Teil 1 ..... A-I

Anlagen, Teil 2 ..... A-II

Anlagen, Teil 3 ..... A-III

Anlagen, Teil 4 ..... A-IV

Selbstständigkeitserklärung VI

Abbildungsverzeichnis

Abbildung 1: Grundlegender Aufbau einer Aminosäure 4

Abbildung 2: Venn Diagramm 5

Abbildung 3: Bildung einer Peptidbindung 5

Abbildung 4: Strukturebenen eines Proteins 8

Abbildung 5: Faltungstrichter 10

Abbildung 6: Darstellung verschiedener Stammbaumtypen 11

Abbildung 7: Erster Schritt bei der Generierung eines UPGMA Baumes 13

Abbildung 8: Zweiter Schritt bei der Generierung eines UPGMA Baumes 13

Abbildung 9: Finaler UPGMA Baum 14

Abbildung 10: Neighbor-Joining Baum 18

Abbildung 11: Beispiel für mögliche Maximum Likelihood Bäume 20

Abbildung 12: Bezeichnung eines Maximum Likelihood Baum 2 20

Abbildung 13: Schematische Darstellung der Bildungsformel eines ML Baumes 23

Abbildung 14: Multiples Sequenzalignment verschiedener BRCT-Domänen 26

Abbildung 15: Struktur der BRCT-Domäne aus TDT 26

Abbildung 16: Phosphorbindungstasche von BRCA1 und RFC 30

Abbildung 17: Genau Betrachtung der Phosphorbindungstasche von RFC 31

Abbildung 18: Struktur des X1BRCTb Homodimeres 33

Abbildung 19: Struktur der homodimeren Bindung von L3BRCT 35

Abbildung 20: Tetramerkomplex aus Monomeren von X1BRCT und L3BRCT 36

Abbildung 21: Komplex aus XRCC4 und DNA Ligase IV..... 39

Abbildung 22: Schematische Darstellung der Abfolge der BRCT-Domänen in TopBP1 40

Abbildung 23: Gegenüberstellung der Phosphorpeptid Bindetasche der siebten und der 0/1/2 Triple BRCT Domäne aus TopBP1 und der BRCT-Domäne aus RFC..... 42

Abbildung 24: Darstellung des evolutionären Ablaufes der BRCT-Domäne 46

Abbildung 25: Die 8Å-Umgebung um His114 des menschlichen Angiogenius 49

Abbildung 26: Gegenüberstellung der energetischen Unterschiede zwischen BARD1-N, BRCA1-N und PTIP 2/3-N..... 68

Abbildung 27: Strukturalignment von BARD1-N, BRCA1-N und MCPH1 2/3-N..... 69

Abbildung 28: Vergleich der Energien an der Position des β2-Sheets in BRCA1-N und BRCA1-C..... 71

Abbildung 29: Gegenüberstellung der Energien des GG-Loops aus vier verschiedenen BRCT-Domänen 75

Abbildung 30: Vergleich der räumlichen Ausrichtung der α2-Helix 76

Abbildung 31: Vergleich der Energien der Bindungstasche einer N-terminalen d1 Domäne mit einer s2 Domäne..... 79

Abbildung 32: Vergleich der Energien der Bindungstasche einer N-terminalen d1 Domäne mit einer s1 Domäne..... 82

Abbildung 33: Schematische Darstellung des neuen evolutionären Verlaufes 84

Tabellenverzeichnis

Tabelle 1: Auflistung aller verwendeten Proteine in denen sich eine oder mehrere BRCT-Domäne befinden.....52

Tabelle 2: Einteilung der BRCT-Domänen in Cluster.....64

1 Einleitung

1.1 Motivation

Die C-terminale Domäne des Breast Cancer Gens 1, oder auch BRCT-Domäne genannt, ist eine in vielen Spezies vorkommende Proteindomäne, welche ein wichtiger Bestandteil des DNA Reparations- und Dedektionsmechanismus ist. Zudem ist sie eine der bisher am besten untersuchten Domänen, da sie, wie der Name schon vermuten lässt, eine entscheidende Rolle bei der Entstehung von Brustkrebs spielt. Somit wurden bisher eine Vielzahl an Studien und Untersuchungen durchgeführt, welche eine gute Beschreibung der funktionellen Eigenschaften, des Vorkommens, der Auswirkungen auf den Organismus und der evolutionären Entwicklung liefern. All diese Ergebnisse wurden mit Hilfe von Sequenzanalysen und Laborexperimenten gewonnen. Jedoch existiert seit einiger Zeit eine Untersuchungsgrundlage, welche mit Hilfe sequenzieller und struktureller Daten und dem Einbezug von Wechselwirkungen, welche innerhalb des Proteins vorkommen, eine genauere Beschreibung und Analyse von Proteinen ermöglicht. Diese neue Grundlage in Form von Protein-Energieprofilen ermöglicht es den gesamten Informationsgehalt, den ein Protein in seiner dreidimensionalen Darstellung besitzt, als zweidimensionales Datenformat abzubilden und zu verarbeiten. Diese neue Grundlage wird momentan in verschiedenen Bereichen der Proteinanalysen auf seine Anwendbarkeit hin untersucht. Da die BRCT-Domäne eine große sequenzielle Vielfalt innerhalb ihrer Proteinfamilie besitzt und diese bisher nur ausführlich auf sequenzieller Grundlage untersucht wurde, sollte nun eine genauere Analyse auf Grundlage der Energieprofile erfolgen, welche evtl. einen besseren Einblick und neue Erkenntnissen über diese liefert. Dabei soll vorrangig die evolutionäre Entwicklung der BRCT-Domäne betrachtet werden und wie sich deren Funktionen im Laufe der Evolution angepasst haben.

1.2 Zielsetzung

Ziel dieser Arbeit war es zum Einen, eine Übersicht bzw. Zusammenfassung der bisher bekannten Funktionen, Eigenschaften und evolutionären Erkenntnissen der BRCT-Domäne zu liefern. Bei anfänglichen Recherchen über die BRCT-Domäne stellte sich nämlich heraus, dass es keine einheitliche und übersichtliche Beschreibung über die Vielzahl an Funktionen, welche die Domäne erfüllt gibt. Somit war es ein Ziel auf Grundlage verschiedenster bisheriger Studien und Untersuchungsergebnissen eine genaue Übersicht über die BRCT-Domäne zu erstellen. Das zweite wichtige Ziel dieser Arbeit sollte es sein, zu überprüfen inwieweit sich mit Hilfe von Energieprofilen phylogenetische Stammbäume und Verwandtschaftsbeziehungen rekonstruieren lassen, da diese bisher noch keine Anwendung auf diesen Forschungsgebiet gefunden haben. Die daraus resultierenden Ergebnisse sollten anschließend auf ihre Glaubwürdigkeit, Plausibilität und Richtigkeit gegenüber bisheriger gewonnener Daten überprüft werden. Das schlussendliche Ziel war es, mit Hilfe von Energieprofilen die bisherigen evolutionären und funktionellen Erkenntnisse der BCRT-Domäne zu überprüfen und evtl. neue zu liefern.

2 Grundlagen Proteine

2.1 Proteine Allgemein

Proteine sind aus den 20 kanonischen Aminosäuren aufgebaute multifunktionale Makromoleküle, welche als Bestandteile jeder Zelle eine Vielzahl an unterschiedlichen Aufgaben und Funktionen besitzen. Diese Aufgaben sind unter anderem der Transport von Ionen, der Metabolismus, das Erkennen von Signalstoffen sowie das Katalysieren von chemischen Reaktionen. Je nach Aufgabe und Funktion kann man die Proteine in verschiedene Gruppen unterteilen. In erster Linie findet eine grobe Unterteilung in globuläre Proteine und Membranproteine statt. Globuläre Proteine sind solche, welche sich frei im Cytoplasma bzw. in Zellorganellen befinden während Membranproteine fest in den Zellwänden oder Wänden von Zellorganellen verankert sind. Eine weitere Unterteilung kann anschließend anhand ihrer Funktion vorgenommen werden, z.B. in Enzyme, Transportproteine, Strukturproteine, kontraktile Proteine, Immunglobuline, Rezeptorproteine und Zellerkennungsproteine. [1]

Wichtig für die Funktionen der Proteine ist deren dreidimensionale Struktur sowie die chemisch-physikalischen Eigenschaften, welche durch die verwendeten Aminosäuren in der Sequenz bestimmt werden. Durch Mutationen in den Aminosäuresequenzen, welche im Laufe der Evolution auftreten, findet eine ständige Veränderung der Strukturen und Funktionen statt. Durch diese Mutationen werden bestehende Funktionen entweder optimiert, gehen verloren oder werden durch neue Funktionen ersetzt. Die Aminosäuresequenz ist auf der DNA kodiert. Dabei kodieren immer drei aufeinander folgende Basen der DNA für eine Aminosäure. Der Grund für diesen Tripletcode ist die Tatsache, dass es 20 verschiedene Aminosäuren gibt, aber nur vier verschiedene Basen (Adenin, Guanin, Cytosin und Thymin) welche diese kodieren können. Bei einer Kodierung durch drei Basen ist es jedoch möglich für 64 verschiedene Aminosäuren zu kodieren, somit besitzen fast alle Aminosäuren mehrere Basenkombinationen. Die Übersetzung der Basenkodierung findet durch die Aminoacetyl-tRNA-Synthease statt. Dieses Protein belädt die tRNA je nach Kodierung mit der spezifisch passenden Aminosäure. Das Aneinanderketten der Aminosäuren erfolgt anschließend durch das Ribosom, wo anhand einer mRNA, welche die Sequenzinformationen der DNA besitzt, die Aminosäuren in der angegebenen Reihenfolge miteinander durch eine Peptidbindung verbunden werden. [1][4]

2.2 Aufbau und Struktur von Proteinen

Wie bereits erwähnt sind Proteine hauptsächlich aus den 20 kanonischen Aminosäuren aufgebaut. Jede Aminosäure besitzt an ihrem einen Ende eine Aminogruppe ($-NH_2$) und an dem anderen eine Carboxylgruppe (-COOH). Diese beiden Gruppen sind über ein zentrales C-Atom, dem sogenannten α-C-Atom, miteinander verbunden und bilden das Rückgrat der Aminosäure. Des Weiteren besitzt jede Aminosäure einen individuellen Rest, welcher die spezielle Seitenkette darstellt und ebenfalls am α-C-Atom befestigt ist (siehe Abbildung 1). Diese individuelle Seitenkette bestimmt die physikochemischen Eigenschaften (z.B. Hydrophobizität, Polarität, Säure-Basen-Verhalten) einer jeden Aminosäure. Die Darstellung der verschiedenen Eigenschaften der Aminosäuren wird häufig in der Form eines Venn Diagrammes vorgenommen, so wie es in Abbildung 2 zu erkennen ist. [2] [3]

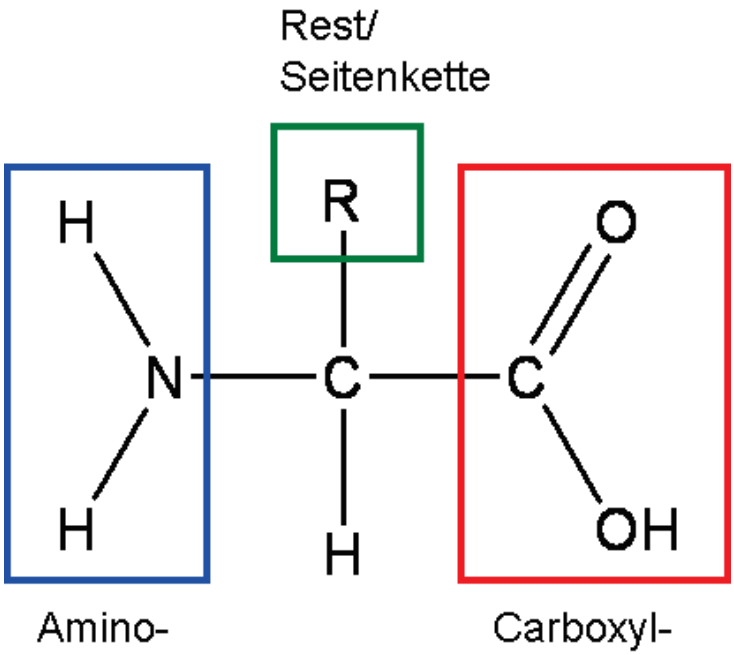


Abbildung 1: Grundlegender Aufbau einer Aminosäure. Rot umrandet ist die Carboxylgruppe, blau die Aminogruppe. Der grüne Kasten repräsentiert einen beliebigen Rest, der je nach Aminosäure unterschiedlich ist. In der Mitte befindet sich das α-C-Atom. [Quelle: http://daten.didaktikchemie.unibayreuth.de]

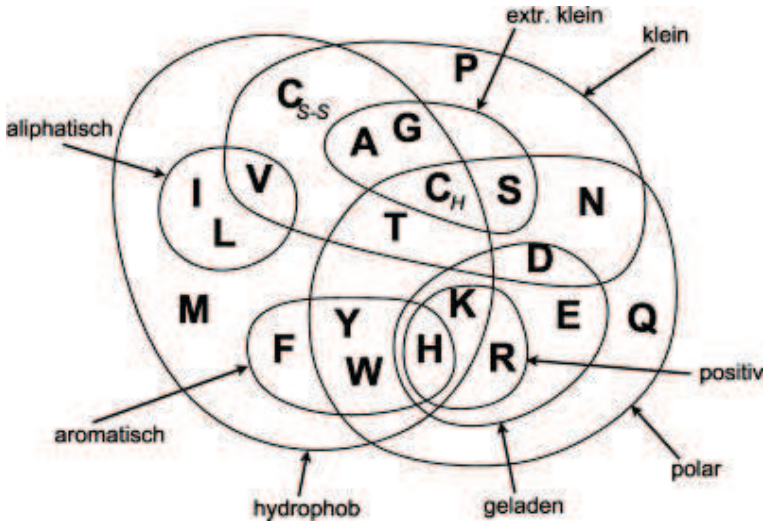


Abbildung 2: Darstellung der Eigenschaften der 20 kanonischen Aminosäuren als Venn Diagramm. Gleiche Eigenschaften werden in Gruppen zusammengefasst, dabei können sich Eigenschaften überschneiden. Bsp.: Aliphatisch = I, L, V. bzw. V = aliphatisch, hydrophob, winzig. [Quelle: http://www-lehre.img.bio.uni-goettingen.de]

Die Bildung von Proteinen erfolgt nun dadurch, dass sich immer zwei Aminosäuren über ihre Amino- und Carboxylgruppe durch Ausbildung einer sogenannten Peptidbindung miteinander verknüpfen und sich dadurch eine lange Kette (Polypeptiedkette) bildet. Diese Kette faltet sich dann durch Einwirkung von chemischen Wechselwirkungen und anderen Faktoren zu einer dreidimensionalen Struktur. Dabei ist wichtig sich zu merken, dass die Biosynthese von Proteinen immer vom Amino- zum Carboxylterminus erfolgt, so dass der entstehende Polypeptidstrang eine Direktionalität erhält. Entsprechend werden Aminosäuresequenzen auch immer in dieser Orientierung angegeben. [1]

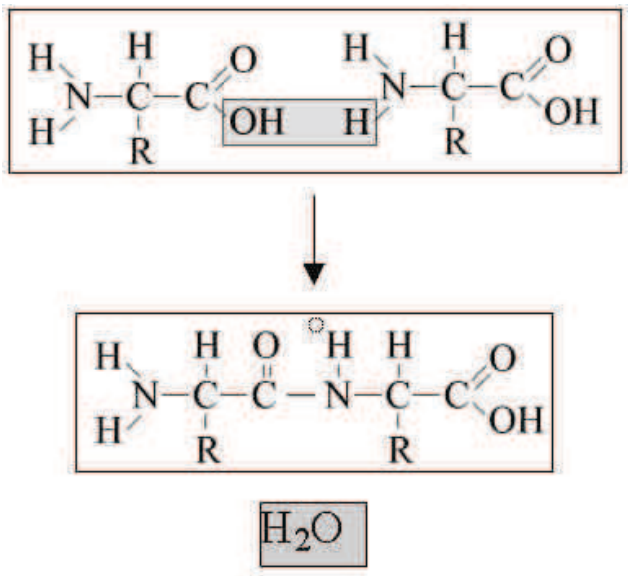


Abbildung 3: Ein Beispiel für das Ausbilden einer Peptidbindung. Durch Abspalten von Wasser bildet sich zwischen der Carboxylgruppe und der Aminogruppe eine Peptidbindung aus. [Quelle: http://www.u-helmich.de/bio/stw/biokatalyse/katalyse02.html]

Um den strukturellen Aufbau der Proteine besser charakterisieren zu können, wurden die Bezeichnungen Primär-, Sekundär-, Tertiär-, und Quatärstruktur eingeführt. Als Primärstruktur bezeichnet man dabei nichts weiter als die Darstellung der Abfolge der einzelnen Aminosäuren in einer Sequenz. Bei der Sekundärstruktur wird zusätzlich zu der Abfolge der Aminosäuren noch die Art und Weise der Kettenfaltung dargestellt, welche durch das Ausbilden von H-Brücken zwischen einigen Aminosäuren zustande kommen. Dabei unterscheidet man zwei wesentliche Strukturen und zwar die α-Helices und die β-Sheets (Faltblätter). In der Tertiärstruktur betrachtet man dann diese Strukturen in ihrer räumlichen Ausrichtung. So liefert sie nicht nur Angaben über die Molekülgestalt, sondern auch über die räumliche Anordnung reaktiver Aminosäurereste, z.B. im aktiven Zentrum von Enzymen oder im Antigenbindungsort von Antikörpern. Die Stabilisierung der Tertiärstruktur wird von einer Vielzahl von Bindungen und Wechselwirkungen zwischen den Aminosäuren gewährleistet. Die Quartiärstruktur als letzte Charakterisierung beschreibt die Interaktion bzw. das Aneinanderbinden von mehreren Proteinen miteinander. Durch dieses Aneinanderbinden bzw. die Ausbildung von Multiproteinkomplexen wird die Effizienz der einzelnen Proteine erhöht. [4]

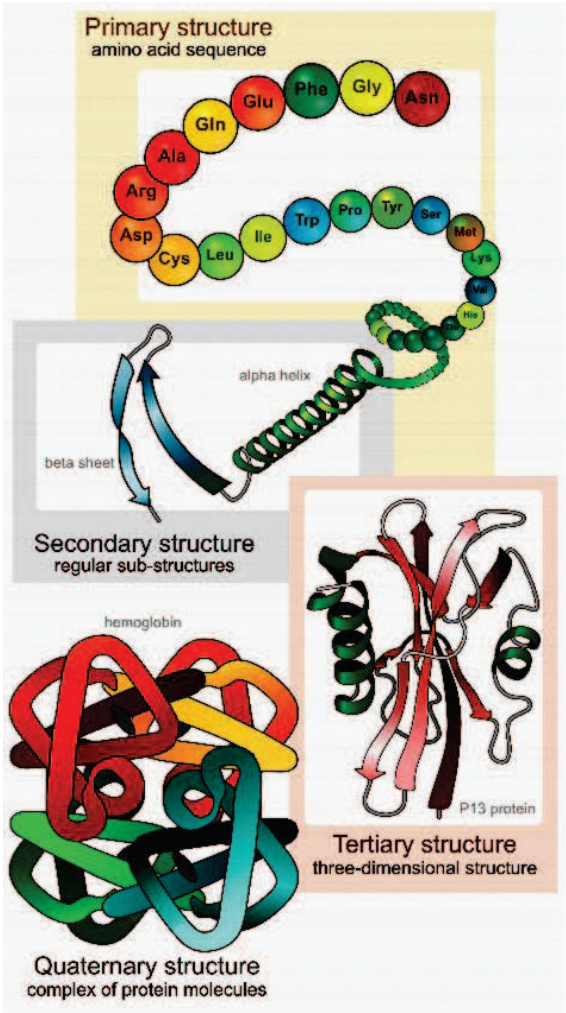


Abbildung 4: Darstellung der Strukturebenen des Proteinaufbaus [Quelle: http://www.walter-w-schuler.com2.html]

2.3 Proteinfaltung

Unter Proteinfaltung versteht man den Übergang von der linearen und ungefalteten Aminosäurekette hin zu einer nativen funktionsfähigen Konformation. Dabei kommt es immer zur Ausbildung der gleichen Tertiärstruktur. Wie dies möglich ist, da es mehr als nur einen stabilen Zustand der Tertiärstruktur pro Protein gibt und wie der genaue Vorgang der Proteinfaltung stattfindet ist momentan jedoch noch ungeklärt.

Die Proteinfaltung in die jeweils entsprechende Konformation erfolgt innerhalb von Mikrosekunden, jedoch existiert rein theoretisch eine unfassbar große Menge an weiteren Konformationen, welche sich aus der zufälligen Faltung der Aminosäurekette ergeben könnten. Dieser Sachverhalt wurde 1968 zum ersten Mal von Cyrus Levinthal formuliert und ist seither als Levinthal-Paradoxon bekannt. Laut diesem Paradoxon kann ein Protein mit *k* Aminosäuren und *n* möglichen Konformationenszuständen n^k mögliche Tertiärstrukturen ausbilden. An einem Beispiel erklärt würde dies bedeuten, dass ein Protein welches aus 150 Aminosäuren besteht und nur 2 mögliche Konformationen besitzt, 2^{150} mögliche Tertiärstrukturen einnehmen kann. Wenn man jetzt noch die Dauer einer Konformationsänderung von etwa 10^{-18}s berücksichtigt, ergibt sich daraus ein Zeitumfang von: $2^{150} \cdot 10^{-18}$s = $1{,}4 \cdot 10^{32}$ s = $4{,}6 \cdot 10^{24}$ Jahren, welche nötig sind um alle Konformationen in diesem Beispiel durchzugehen. Diese Zeitspanne entspricht jedoch einem Vielfachen dem Alter der Erde, welche im Vergleich dazu das Alter von $4{,}55 \cdot 10^9$ Jahren aufweist. [2] [3]

Anhand dieser enorm hohen Zahl kann man davon ausgehen, dass die Proteinfaltung nicht zufällig stattfindet, sondern bestimmte Faltungspfade existieren müssen, welche dafür sorgen, dass sich die Aminosäurekette immer in ihre gleiche Form faltet. Einige Modelle, welche versuchen die Proteinfaltung zu beschreiben, wären das Gerüstmodell, das Modell des hydrophoben Kollaps und das des Faltungstrichters. Bei dem Gerüstmodell oder auch „framework-model“ genannt geht man davon aus, dass die Faltung der dreidimensionale funktionellen Struktur damit beginnt, dass sich zuerst die Sekundärstrukturelemente (α-Helix, β-Sheet oder random coil-Anteile) ausbilden und somit als Gerüst dienen, bevor sie damit beginnen, sich im Raum als Tertiärstruktur auszurichten. Jedoch ist dieses Modell nicht allgemein zutreffend bzw. kann nicht so einfach übertragen werden, da Beobachtungen zeigten, dass viele Sekundärstrukturelemente innerhalb von gefalteten Proteinen nicht mehr die gleiche Faltung besitzen wie am Anfang. [5] [6]

Bei der Theorie des hydrophoben Kollaps geht man davon aus, dass die treibende Kraft hinter diesem Faltungsweg die Freisetzung von komplexierten Wassermolekülen ist. Hierbei wird der Entropiegewinn des Systems durch das freigesetzte Wasser als maßgeblich

treibende Kraft für die Proteinfaltung angesehen. Jedoch müsste die Aminosäurekette dazu in einem völlig entfaltetem Zustand, also in einem Zustand in dem alle Aminosäuren sich möglichst unabhängig voneinander verhalten können vorliegen damit diese Kraft des hydrophoben Kollaps sich voll entfalten kann. Die meisten Proteine weisen jedoch in ihrem ungefalteten Zustand eine schnelle Fluktuation von Sekundärstrukturen auf. Deshalb geht man in der Regel davon aus, dass oft gewisse Minimalstrukturen vorhanden sind und somit die hydrophoben Kräfte allein nicht ausreichen um die Proteinfaltung bestimmen zu können. [5][6]

Da nun die Modelle der Faltung alle irgendwo gewisse Stärken und Schwächen zueinander aufwiesen, vereint man diese zu einem Gesamtmodell, welches als Faltungstrichter bezeichnet wird. Dieses Modell lässt unterschiedlichste Faltungsmöglichkeiten und –wege für ein einzelnes Protein zu und macht Angaben über eine Vielzahl von Faltungsintermedianten (dargestellt als lokale Minima) und einem globalen Minimum, welches die native Struktur darstellt. Das Protein kann sich nach dieser Vorstellung auf dem kinetisch schnellsten Weg oder über mehrere lokale Minima (Intermdeiate) und Maxima (Übergangszustände) auf einem nicht-direkten Weg falten.

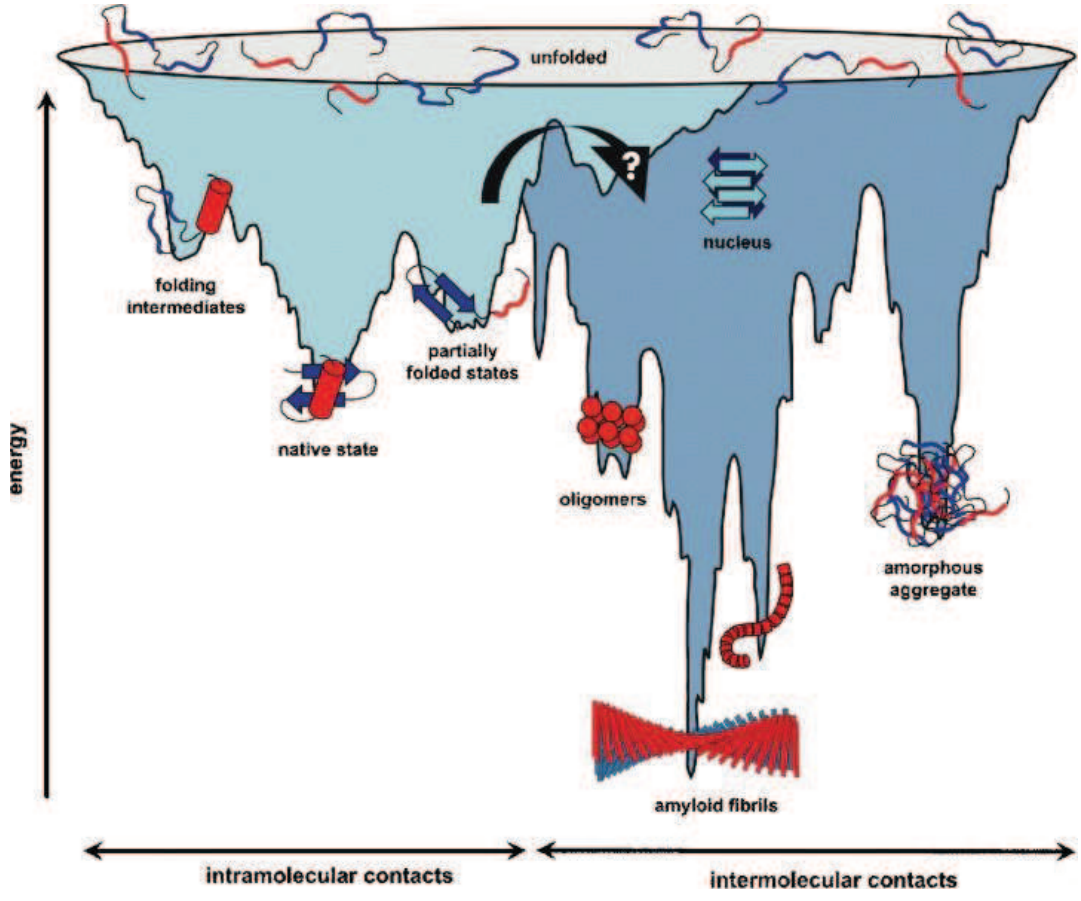


Abbildung 5: Faltungstrichter für Proteine unter Berücksichtigung der Fehlfaltung [6]

3 Grundlagen phylogenetischer Methoden

3.1 Phylogenie Allgemein

Als Phylogenie oder Phylogenese bezeichnet man das Forschungsfeld, welches sich mit der stammesgeschichtlichen Entwicklung der Lebewesen beschäftigt. Dabei verfolgt man das Ziel, die Verwandtschaftsbeziehungen zwischen den biologischen Arten, Populationen, Individuen oder Genen aufzuklären Bei der molekularen Phylogenie handelt es sich um die Beschreibung der Verwandtschaftsbeziehungen auf der molekularen Ebene von Genen und Proteinen. Dabei beobachtete man die Veränderungen von Nukleotidsequenzen, Proteinsequenzen, Proteinstrukturen oder auch Proteinenergien. Dies bietet gegenüber der Betrachtung von morphologischen und physiologischen Eigenschaften große Vorteile. Zum einen werden Sequenzen direkt vererbt und unterliegen somit keinen Umwelteinflüssen. Des Weiteren erlauben sie Vergleiche über große Distanzen und Spezies hinweg (Vergleich von Tieren, Pflanzen, Pilzen, Bakterien,... miteinander möglich) und sind weitgehend frei von Interpretationseinflüssen (z.B. „dick“, „etwas abgeflacht“, „rundlich“ etc.). Dank einer Vielzahl von Onlinedatenbanken und Sequenziermethoden sind Sequenzdaten mittlerweile in großen Mengen relativ schnell und kostengünstig zu erhalten. Dennoch müssen auf molekularen Daten basierende Stammbäume nicht zwingend auch die richtige Evolution darstellen. [7][8]

Die Beschreibung der Verwandtschaftsbeziehungen findet meist in Form eines Stammbaumes statt. Diese Stammbäume können entweder die Form eines gewurzelten (rooted tree) oder ungewurzelten (unrooted tree) Baumes haben. Bei einem gewurzelten Baum gehen alle Nachkommen auf einen gemeinsamen Urahn zurück, von welchem sich alle Spezies aus entwickelt haben. Somit zeigt ein solcher Baum die Annahme über die Richtung der Evolution und deren historischen Verlauf. Ein ungewurzelter Baum dagegen gibt nur Auskunft über die Verwandtschaft der Spezies zueinander. Im Allgemeinen bezeichnet man das Erscheinungsbild des Stammbaumes als Topologie. Die Topologie spiegelt somit die evolutionären Zusammenhänge der einzelnen Taxa wieder. Als Taxon bezeichnet man wiederum eine als systematische Einheit benannte Gruppe von Lebewesen. In der molekularen Phylogenie sind dies z.B. die Sequenzen eines Gens oder Proteins. Eine Operative Taxonomische Einheit („operational taxonomic unit“ = OTU) bezeichnet in einem Stammbaum die außenstehenden sichtbaren Spezies, Gene oder ähnliches. Diese OUTs sind durch Äste (branch) über Knotenpunkte (node) miteinander verbunden. Die

internen Knotenpunkte, von welchen mindestens drei Äste ausgehen werden auch als Hypothetische Taxonomische Einheiten („hypothetical taxonomic unit = HTU“) bezeichnet. [9][10][11][8]

Allgemein gibt es drei Typen von Stammbäumen, die es zu unterscheiden gilt. Bei einem Kladogramm besitzen die Längen der terminalen und inneren Äste keinerlei Bedeutung. Lediglich die Topologie und das Verzweigungsmuster sind ausschlaggebend. Bei einem Phylogramm (oder auch metrischer Stammbaum genannt) wiederum spielt die Länge der einzelnen Äste eine wichtige Rolle. So entspricht die Astlänge z.B. der Anzahl der beobachteten Merkmalsaustausche im Laufe der Evolution oder einer molekulargenetischen Distanz, welche nach einem bestimmten Maß gewählt wurde. Das Phylogramm ist die am häufigsten genutzte Form von Stammbäumen, da es, anders als das Kladogramm, zusätzlich ein Größenmaß für die evolutionären Veränderungen in seinen Astlängen liefert. Der letzte Typ von Stammbäumen sind sogenannte Chronogramme. In diesen repräsentieren die Länge der Äste nicht die evolutionären Veränderungen wie in einem Phylogramm sondern die Zeit, in welcher sich die OTUs entwickelt haben. In Abbildung 6 sind die verschiedenen Formen und Typen von Stammbäumen zu erkennen. [9][10][11][8]

Um nun aus molekularbiologischen Daten, von Genen oder Proteinen auf evolutionäre Zusammenhänge schließen zu können und um diese in Stammbäumen darzustellen zu können, wurde eine Vielzahl von unterschiedlichen Analysemethoden und Algorithmen entwickelt. Diese Methoden lassen sich in zwei grundlegende Kategorien einteilen. Zum einen in Distanz Methoden und zum anderen in Charakter Methoden. Molekularbiologische Daten von DNA und Proteinsequenzen sind grundlegend Charakterdaten, welche jedoch in Distanzdaten umgewandelt werden können. Der Grund für die Umwandlung in Distanzdaten ist oft die einfachere Erfassbarkeit und Verarbeitung dieser. So können Charaktermethoden auf enorm größere Datenmengen angewandt werden, als es mit Charaktermethoden möglich ist, da deren Berechnungsalgorithmen einfacher arbeiten. Jedoch wird dabei ein Teil der phylogenetischen Informationen „verschenkt“, was dazu führt, dass sie die evolutionären Zusammenhänge ungenauer darstellen. Jedoch besitzen alle phylogenetischen Methoden, welche man auf molekularbiologische Daten anwenden kann, den gleichen Ausgangspunkt. Dieser Ausgangspunkt ist ein MSA, aus welchem Substitutionsmatrizen generiert werden. Charaktermethoden beziehen ihre Informationen direkt aus diesen Substitutionsmatrizen während bei den Distanzmethoden die Informationen der Substitutionsmatrizen in Distanzmaße umgerechnet und in eine Distanzmatrize überführt werden. [7][9][10]

Die am meisten genutzten Distanzmethoden sind die UPGMA und Neighbor-Joining Methode. Zu den Charaktermethoden zählen z.B. die Maximum-Parsimony, Maximum-Likelihood und Bayes Methode. Im weiteren Verlauf dieser Arbeiten wurden hauptsächlich die UPGMA, Neighbor-Joining und Maximum-Likelihood Methode benutzt. Die genaue Funktionsweise dieser Methoden und deren Algorithmen soll nun im weiteren Verlauf näher erläutert werden.

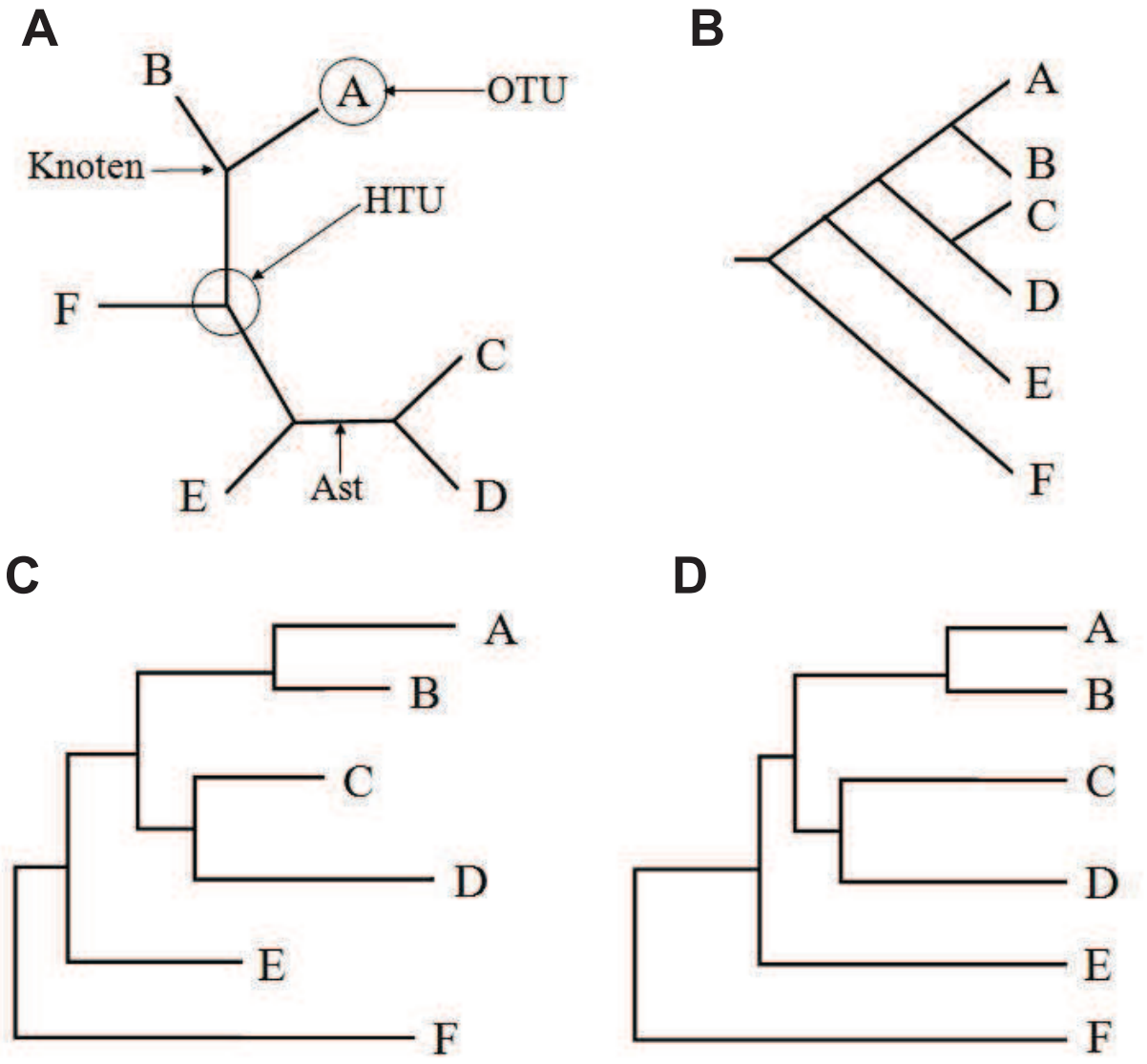


Abbildung 6: Darstellung verschiedener Stammbaumtypen und –formen. A: Schematische Darstellung des Aufbaus eines Stammbaumes, welcher gleichzeitig einen ungewurzelten Baum darstellt. B: Abbildung eines Kladogramms als gewurzelter Baum. C: Darstellung eines Phylogramms. Die Länge der Äste repräsentiert eine vorher definierte evolutionäre Distanz. D: Beispiel eines Chronogramms.

3.2 Die UPGMA Methode

Die Unweighted Pair Group Method with Arithmetic mean oder auch kurz UPGMA genannt, ist eine auf Distanzmatrizen beruhende bottom-up clustering Methode, welche zur Rekonstruktion von phylogenetischen Bäumen genutzt wird. Anders als die anderen phylogenetischen Methoden, wie z.B. die artverwandte Neighbor-Joining Methode, basiert UPGMA auf der Annahme einer molekularen Uhr. Dies bedeutet, dass sie eine konstante Evolutionsgeschwindigkeit der einzelnen Spezies annimmt, wodurch alle Taxa mit derselben konstanten Rate evolvieren und somit keine Aussage über die Geschwindigkeit, mit der sich die Evolution vollzogen hat gemacht werden kann. [9][10][12]

Der Ausgangspunkt für den Algorithmus ist eine Distanzmatrix, welche die paarweisen Distanzen der einzelnen Objekte (entspricht den einzelnen Spezies, Sequenzen, ... aus dem MSA) enthält. Diese Distanzmatrix wird aus den einzelnen Substitutionsmatrizen berechnet, welche wiederrum aus dem ausgehenden MSA generiert wurden. Die Distanzmaße müssen jedoch die Eigenschaft der Ultrametrik, d.h. die Erfüllung der Dreiecksungleichung aufweisen, ansonsten ist die Anwendung des UPGMA Algorithmus nicht möglich. [9][10][12]

Zu Beginn der Berechnung ist jedes Objekt ein eigener Cluster. Im jeden Schritt werden die beiden Cluster, welche das geringste Distanzmaß zueinander aufweisen in einem neuen Cluster zusammengefasst und anschließend die Distanzmatrix neu berechnet. Die Distanz zwischen den beiden Clustern entspricht dann dem Mittelwert der paarweisen Distanzen der Objekte in den Clustern. Wenn man nun z.B. von einer Distanzmatrix mit den Objekten A, B, C und D (= Cluster A, B, C und D) ausgeht, welche die dazugehörigen Distanzmaße d_{AB} bis d_{CD} enthält, so berechnet sich ein UPGMA Baum wie folgt:

Ausgangspunkt ist folgende Distanzmatrix:

	A	**B**	**C**
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

Es wird angenommen, dass d_{AB} das kleinste Distanzmaß ist. Somit werden nun die Cluster A und B zu einem neuen Cluster (AB) zusammengefasst und die Distanzen zu diesem neuen Cluster wie folgt berechnet:

$$d_{(AB)C} = \frac{d_{AC} + d_{BC}}{2} \quad \text{bzw.} \quad d_{(AB)D} = \frac{d_{AD} + d_{BD}}{2}$$

Somit ergibt sich dann folgende neue Distanzmatrix:

	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

Der nun aus dem ersten Schritt resultierende Baum verbindet die beiden Cluster A und B über Taxa miteinander, dessen Distanz dem Mittelwert der Distanzen der beiden Objekte entspricht.

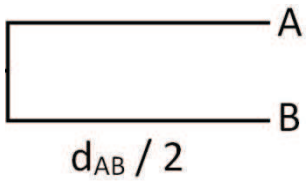


Abbildung 7: Der erste Baum des Beispiels, welcher entsteht wenn die Objekte A und B über einen Ast miteinander verbunden werden. Der Abstand zwischen beiden Objekten wird mit dem Maß d_{AB} / 2 angegeben.

Aus der neu berechneten Distanzmatrix werden nun wiederum die beiden Cluster mit dem geringsten Distanzmaß zueinander zu einem neuen Cluster zusammen geführt und in dem Baum durch Taxa miteinander verbunden. Angenommen in diesem Beispiel wäre dies nun $d_{(AB)C}$, so wird nun C der OTU (AB) zugeordnet mit der Astlänge $d_{(AB)C}$ /2.

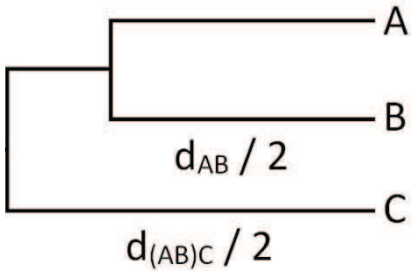


Abbildung 8: Der zweite Baum des Beispiels, welcher entsteht wenn zu den Objekten A und B das Objekt C, über einen Knoten mit dem Abstand $d_{(AB)C}$ / 2 angefügt wird.

Zum Schluss wird noch die Distanz des neuen Clusters (ABC) zu dem Cluster D ermittelt und dieser dann mit dem Baum verbunden.

$$d_{(ABC)D} = \left(\frac{d_{AD} + d_{BD} + d_{CD}}{3}\right) \div 2$$

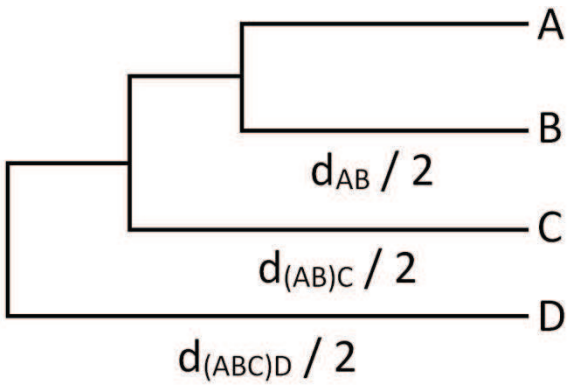


Abbildung 8: Der dritte und letzte Baum des Beispiels, an welchen das letzte Objekt D mit dem Abstand $d_{(ABC)D}$ / 2 an den Baum aus Abbildung 9 angefügt wurde.

Der am Ende resultierende Baum stellt die evolutionären Zusammenhänge jedoch nur ungenau dar. Dies liegt zu einen daran, dass UPGMA eine auf Distanzen basierende Methode ist und diese im Gegensatz zu Charakter Methoden, wie z.B. Maximum Likelihood auf schlichteren bzw. einfacheren mathematischen Algorithmen beruhen. Von daher weicht die Topologie der UPGMA Bäume oft von denen der Charakter Methoden ab und dient von daher nur zu einer schnellen und groben Übersicht der evolutionären Zusammenhänge.

3.3 Die Neighbor-Joining Methode

Der Neighbor-Joining-Algorithmus (NJ) ist ein 1987 von Naruya Saiou und Masatoshi Nei entwickeltes mathematisches Verfahren zur Rekonstruktion von phylogenetischen Bäumen. Wie auch der UPGMA-Algorithmus ist es eine auf Distanzmatrizen beruhende bottom-up clustering Methode, welche jedoch im Gegensatz zu UPGMA keine konstante Evolutionsrate annimmt, sondern die jeweilige individuelle Evolutionsgeschwindigkeit der Spezies bzw. OTU berücksichtigt. Dies zeigt sich dadurch, dass ein Taxon, welches von allen anderen Taxa im Baum weit entfernt ist, eine beschleunigte Evolution vollzogen hat.

Des Weiteren basiert der Algorithmus auf der Annahme der minimalen Evolution, was bedeutet, dass der Stammbaum, dessen Summe an Astlängen am kleinsten ist die Evolution am richtigsten darstellt. [9][10][12][13]

Ausgangspunkt eines NJ Baumes ist ein sternenförmiger Baum, in dem alle Taxa mit der gleichen Länge über ein gemeinsames Zentrum miteinander verbunden sind. Von diesem Baum aus werden anschließend paarweise die Objekte mit der geringsten Distanz ausgewählt und zu einem Ast des Baumes vereint. Diese Abstände werden aus einer Zwischenmatrix entnommen, welche aus der Distanzmatrix errechnet wird. Die Werte der Distanzmatrix wiederum erschließen sich wie bei dem UPGMA Algorithmus aus den Substitutionsmatrizen des MSA. Anschließend, nach der Vereinigung der zwei Objekte über einen Ast, wird die Distanzmatrix neu berechnet. Nun erfolgt erneut ein Zusammenschluss der beiden Objekte mit der geringsten Distanz und eine darauffolgende Neuberechnung der Distanzmatrix. Dieser Vorgang wird so lang wiederholt, bis sich alle Taxa in dem Baum eingefügt haben und sich die Sternenstruktur des Ausgangsbaumes aufgelöst hat. Eine genauere Beschreibung des Vorgehens und der dazugehörigen Berechnungen soll nun anhand des folgenden Beispiels näher erläutert werden.

Ausgangspunkt ist folgende Distanzmatrix, mit folgenden hypothetischen Werten:

	A	B	C	D
A	0	d_{AB}	d_{AC}	d_{AD}
B	d_{AB}	0	d_{BC}	d_{BD}
C	d_{AC}	d_{BC}	0	d_{CD}
D	d_{AD}	d_{BD}	d_{CD}	0

→

	A	B	C	D
A	0	3	14	12
B	3	0	13	11
C	14	13	0	4
D	12	11	4	0

Im ersten Schritt müssen nun zunächst die durchschnittlichen Distanzen von jedem Taxon zu jedem anderen Taxon berechnet werden. Dies erfolgt mit folgender Formel:

$$r_i = \frac{1}{N-2} \sum_{j=1}^{N} d_{i,j}$$

Dabei steht N für die Anzahl der Taxa und $d_{i,j}$ für die Distanz zwischen zwei beliebigen Objekten. Die durchschnittliche Distanz r_i, oder auch Netto-Divergenz genannt wird somit

für jedes Objekte der Distanzmatrix berechnet. In diesem Beispiel sind dies die Objekte A, B, C und D deren entsprechender r_A, r_B, r_C und r_D Wert wie folgt sind:

$$r_A = \frac{0 + 3 + 14 + 12}{4 - 2} = 14{,}5 \qquad r_B = \frac{3 + 0 + 13 + 11}{4 - 2} = 13{,}5$$

$$r_C = \frac{14 + 13 + 0 + 4}{4 - 2} = 15{,}5 \qquad r_D = \frac{12 + 11 + 4 + 0}{4 - 2} = 14{,}5$$

Betrachtet man nun diese Werte, so lässt sich daraus schließen, dass das Objekt C die größte Evolutionsgeschwindigkeit durchlebt hat, da dessen Netto-Divergenz die größte ist. Im nächsten Schritt erfolgt nun die Berechnung einer Zwischenmatrix (M). Die Werte dieser Zwischenmatrix dienen später dazu, die zwei Objekte mit dem geringsten Wert zu einem Ast des Baumes zu vereinen. Die Berechnung der neuen Distanzwerte erfolgt nach folgender Berechnungsvorschrift:

$$m_{i,j} = d_{i,j} - (r_i + r_j)$$

Somit ergibt sich beispielsweise für den Distanzwert d_{AB} der neue Distanzwert m_{AB} von:

$$m_{AB} = d_{AB} - (r_A + r_B) \quad \rightarrow \quad m_{AB} = 3 - (14{,}5 + 13{,}5) = -25$$

Nach der Berechnung aller neuen Distanzwerte $m_{i,j}$ ergebe sich nun folgende Zwischenmatrix:

M	A	B	C	D
A	0	-25	-16	-16
B	-25	0	-16	-16
C	-16	-16	0	-25
D	-16	-16	-25	0

Aus dieser Zwischenmatrix werden nun die zwei Objekte mit dem kleinsten Distanzwert gesucht und zu einem neuen Teilbaum u = (i,j) zusammengesetzt. In diesem Beispiel ergeben sich nun zwei Möglichkeiten für einen neuen Teilbaum. Zum einen besteht die Möglichkeit, dass die Objekte A und B einen Teilbaum bilden oder D und C. Welche der Möglichkeiten nun genommen werden soll spielt keine Rolle, von daher soll an dieser Stelle der Teilbaum u = (AB) gewählt werden. Somit sind nun die Objekte A und B über

den Knoten u miteinander verbunden. Die Länge der Taxa von A und B zu dem neuen Knoten u berechnet sich nun wie folgt:

$$v_{i,u} = \frac{d_{i,j} + r_i + r_j}{2} \quad \rightarrow \quad v_{A(AB)} = \frac{d_{AB} + r_A - r_B}{2} \quad \rightarrow \quad v_{A(AB)} = \frac{3 + 14{,}5 - 13{,}5}{2} = 2$$

$$v_{j,u} = d_{i,j} - v_{i,u} \quad \rightarrow \quad v_{B(AB)} = d_{AB} - v_{A(AB)} \quad \rightarrow \quad v_{B(AB)} = 3 - 2 = 1$$

Nachdem nun zwei Objekte zu einem neuen Teilbaum zusammengefügt wurden, wird der neue Eintrag des Teilbaumes u = (i,j) = (AB) an die ursprüngliche Distanzmatrix angefügt. Zu diesem neuen Eintrag werden nun die Distanzen der restlichen Taxa neu berechnet. Die ursprünglichen Einträge i und j werden dann anschließend aus der Matrix gelöscht, da diese zu dem einem neuen Eintrag u zusammengefügt wurden. Davor werden sie jedoch noch benötigt, um die Distanzen von u zu denen von k zu ermitteln.

$$d_{u,k} = \frac{d_{i,k} + d_{j,k} - d_{i,j}}{2}$$

Anhand des Beispiels würden nun die neuen Distanzen von dem Teilbaum u = (AB) zu dem noch bestehenden Objekten k = C bzw. k = D wie folgt berechnet:

$$d_{(AB)C} = \frac{d_{AC} + d_{BC} - d_{AB}}{2} \quad \rightarrow \quad d_{(AB)C} = \frac{14 + 13 - 3}{2} = 12$$

$$d_{(AB)D} = \frac{d_{AD} + d_{BD} - d_{AB}}{2} \quad \rightarrow \quad d_{(AB)C} = \frac{12 + 11 - 3}{2} = 10$$

Somit ergibt sich nach der Berechnung dieser neuen Distanzen und der Löschung der alten Distanzen von A und B folgende neue Distanzmatrix:

	A	B	C	D	(AB)
A	0	3	14	12	?
B	3	0	13	11	?
C	14	13	0	4	?
D	12	11	4	0	?
(AB)	?	?	?	?	0

→

	C	D	(AB)
C	0	4	12
D	4	0	10
(AB)	12	10	0

Aus dieser neuen Distanzmatrix werden nun wiederum die Werte für r_i und die Zwischenmatrix $M_{i,j}$ berechnet und der soeben erläuterte Ablauf wiederholt bis nur noch zwei Taxa übrig bleiben, welche dann schlussendlich miteinander verbunden werden und somit einen fertigen NJ-Baum ergeben. Der fertig generierte Baum dieses Beispiels sowie der sternenförmige Ausgangsbaum finden sich in der Abbildung 10.

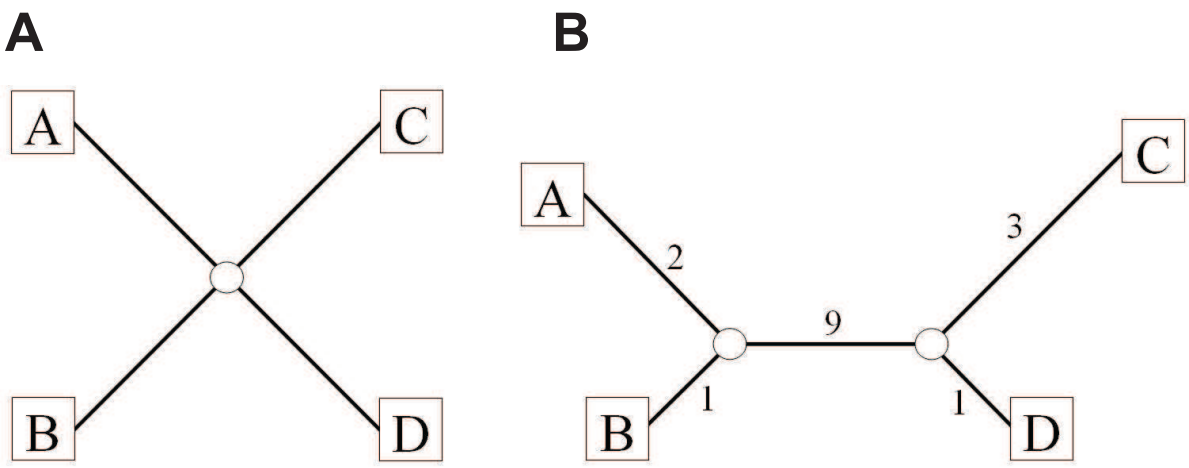


Abbildung 10: A: Sternenförmiger Neighbor-Joning Baum der als Ausgangspunkt für die Berechnung dient. B: Der Ergebnis Baum, welcher am Ende der Berechnungen des im Text erläuterten Beispiels entsteht. Die Zahlen repräsentieren die evolutionären Abstände der OTUs zu dem jeweiligen Knoten.

Das NJ Verfahren bietet gegenüber dem UPGMA Verfahren klare Vorteile, da es zum einen, wie bereits erwähnt, nicht auf der Annahme einer konstanten Evolution beruht, sondern stattdessen die evolutionäre Geschwindigkeit berücksichtigt und diese im Baum wiederspiegelt. Da der NJ Algorithmus, ebenfalls wie der UPGMA Algorithmus auf Distanz Methoden basiert, ist er gegenüber Charakter Methoden ungenauer. Im Gegensatz zu diesen wie z.B. dem Maximum Likelihood Algorithmus berechnet der NJ Algorithmus den Stammbaum schrittweise, was dazu führt, dass während des Verfahrens einige Rechenwege verworfen werden. Bei dem Maximum Likelihood Algorithmus hingegen werden alle möglichen Bäume generiert und am Ende der optimalste ausgewählt. Jedoch wurde in ausführlichen Tests und diversen Untersuchungen festgestellt, dass in den meisten Fällen der entstandene NJ Baum dem Optimum und den Bäumen der Charakter Methoden relativ nahe kommt. Der größte Vorteil, welchen das Verfahren bietet, ist die Geschwindigkeit der Berechnung. Dank dem im Vergleich zu Charakter Methoden einfach gehaltenen Algorithmus, kann das NJ Verfahren auf gewaltige Datenmengen angewandt werden und bietet dort, wo andere phylogenetische Methoden wie z.B. Maximum Likelihood nicht mehr durchführbar sind immer noch die Möglichkeit der Untersuchung evolutionärer Zusammenhänge. [12][13]

3.4 Maximum Likelihood

Die Maximum Likelihood Methode (ML) bezeichnet in der Phylogenie ein statistisches Schätzverfahren zur Ermittlung des wahrscheinlichsten Evolutionsverlaufes. Es ist ein auf Charakter Methoden beruhendes analytisches Verfahren, welches alle möglichen evolutionären Verläufe des zu Grunde legenden Datensatzes betrachtet und am Ende den Verlauf in Form eines Stammbaum ausgibt, welcher die Evolution am wahrscheinlichsten wiederspiegelt. Dieser Stammbaum (ML Baum) repräsentiert die wahrscheinlichste Evolution, welche jedoch nicht zwingend auch die richtige Evolution sein muss. Dennoch findet das ML Verfahren heutzutage die meiste Anwendung unter den phylogenetischen Methoden. Anders als UPGMA und NJ basiert das Verfahren nicht auf Distanzen, sondern bezieht seine Werte direkt aus den Substitutionsmatrizen, welche aus dem MSA generiert werden. Somit wird der Zwischenschritt der Generierung einer Distanzmatrix und der evtl. daraus resultierende Informationsverlust übergangen. Der genaue Ablauf zu Bestimmung eine ML Baumes soll nun anhand folgenden Beispiels näher erläutert werden. [12][14]

Ausgangspunkt für die Ermittlung eines jeden ML Baumes ist ein zugrundeliegendes MSA. In dem jetzt folgenden Beispiel ist dies ein hypothetisches MSA von DNA Nukleotidsequenzen von vier verschiedenen Spezies, genannt Spezi1, Spezi2, Spezi3 und Spezi4.

Position:	1	2	3	4	5	6	7	8	9	...
Spezi1	A	A	C	T	G	T	G	-	C	...
Spezi2	C	A	T	G	G	T	G	-	T	...
Spezi3	C	G	A	-	T	C	A	-	T	...
Spezi4	T	C	-	-	G	C	A	G	C	...

Da nun das als Ausgangspunkt dienende MSA vier Spezies beinhaltet, ergeben sich daraus drei Möglichkeiten der Verwandtschaftsbeziehung, welche zwischen diesen möglich sind. Dies bedeutet, dass es drei mögliche Stammbäume (siehe Abbildung 11) gibt, von denen einer der ML Baum ist. Die Bestimmung, welcher dieser Bäume der ML Baum ist, erfolgt anhand des Likelihood Wertes (L-Wert). Dieser L-Wert wird für jeden der drei möglichen Bäume bestimmt, indem für jeden Baum an jeder Position die Wahrscheinlichkeit ermittelt wird, dass sich eine Aminosäure im Laufe der Evolution durch eine andere austauscht. [12][14]

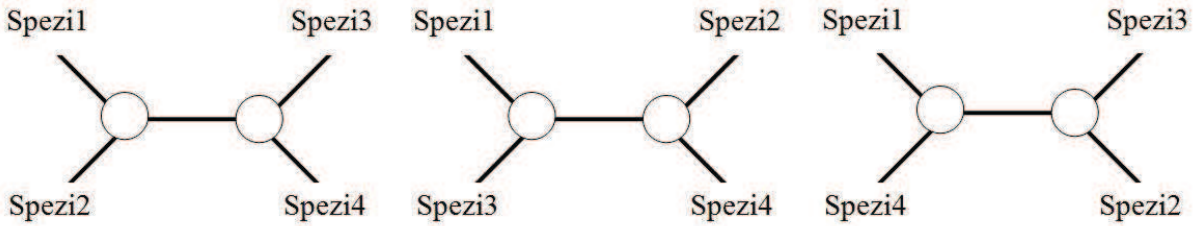


Abbildung 11:: Die drei möglichen Stammbäume, welche laut dem im Text dargestellten MSA möglich sind.

Um dies nun genauer zu erklären, wird zunächst der erste Baum aus Abbildung 11 näher betrachtet. Um für diesen ersten Baum den L-Wert, im weiteren Verlauf L_1 genannt zu ermitteln, müssen zunächst die einzelnen L-Werte aller Positionen dieses ersten Baumes, welche im MSA vorhanden sind, ermittelt werden. Ein jeder L-Wert der jeweiligen Position im MSA ($L_{1\text{-}1}$, $L_{1\text{-}2}$, $L_{1\text{-}3}$ bis $L_{1\text{-}n}$ für die n-Positionen) besteht dabei aus der Summe der Wahrscheinlichkeiten (P), mit welcher die Nukleotide x_{s1}, x_{s2}, x_{s3} und x_{s4} über alle Möglichkeiten (m) zu Nukleotid Y mutieren können. Die Möglichkeiten der Mutation sind je nach Art der Daten des MSA anders gegeben. Im Falle der hier verwendeten Nukleotidsequenz gibt es vier Möglichkeiten des Mutationsverlaufes, nämlich über die vier möglichen Basen der DNA (Adenin (A), Guanin (G), Cytosin (C), Thymin (T)). Im Fall eines MSA, welches aus Proteinequenzen besteht, würden sich 20 Möglichkeiten des Mutationsverlaufes ergeben, da der Austausch einer jeden Aminosäure durch eine jede in Betracht gezogen wird. [12][14]

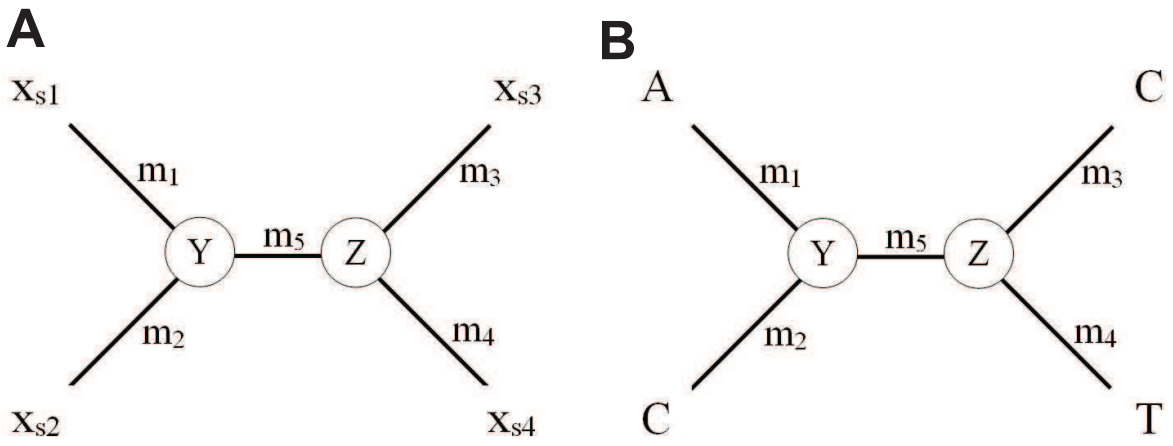


Abbildung 12: A: Schematische Darstellung eines Ausgangsbaumes für vier Spezies. x_{s1} bis x_{s4} stehen für die jeweilige Base oder Aminosäure der Spezies eins bis vier, welche sich an der betrachteten Position im MSA befindet. Y und Z repräsentieren alle möglichen Basen oder Aminosäuren über die x_{s1} zu x_{s2} bzw. x_{s3} zu x_{s4} mutieren können. m_1 bis m_5 geben die Möglichkeit bzw. Wahrscheinlichkeit an, mit welcher die Mutation erfolgt. B: Der erste mögliche Baum für Position 1 des MSA von welchem der $L_{1\text{-}1}$-Wert bestimmt werden soll.

Durch einsetzen der vier möglichen Basen der DNA für Y und X in Baum B aus Abbildung 12 ergeben sich 16 mögliche evolutionäre Verläufe bzw. Bäume für diese Position. Im Falle einer Proteinsequenz würden sich durch einsetzen der 20 Aminosäuren insgesamt 400 mögliche Bäume an dieser ersten Position ergeben. Für die Bestimmung des $L_{1\text{-}1}$-Wertes dieses Baumes müssen nun zunächst die P-Werte der 16 möglichen Mutationsverläufe ermittelt werden. Jeder P-Wert in diesem Beispiel berechnet sich in seiner allgemeinen Form wie folgt:

$$P = (x_{s1}/Y, m_1) \cdot (x_{s2}/Y, m_2) \cdot (Y/Z, m_5) \cdot (x_{s3}/Z, m_3) \cdot (x_{s4}/Z, m_4)$$

Dabei gibt $(x_{s1}/Y, m_1)$ die Teilwahrscheinlichkeit an, mit welcher die Base x_{s1} zu der Base Y mit der Möglichkeit/Wahrscheinlichkeit m_1 mutiert. Die Werte für m_1 bis m_5 werden dabei einer Substitutionsmatrize entnommen, welche aus dem MSA generiert wurde. Die Werte dieser Substitutionsmatrize geben an, mit welcher über dem Zufall liegenden Wahrscheinlichkeit, eine Base oder Aminosäure im Laufe der Evolution zu einer anderen mutiert. Ein Beispiel für solch eine Substitutionsmatrize soll folgende, mit hypothetischen Mutationswahrscheinlichkeiten für die vier Basen der DNA gefüllte Matrix sein:

	A	T	C	G
A	1	0,1	0,1	0,1
T	0,1	1	0,1	0,1
C	0,1	0,1	1	0,1
G	0,1	0,1	0,1	1

Die P-Werte (P1 bis P16) für die Bestimmung des $L_{1\text{-}1}$-Wertes lassen sich nun mit folgender Formel berechnen:

$$P1 \text{ bis } P16 = (A/Y, m_1) \cdot (C/Y, m_2) \cdot (Y/Z, m_5) \cdot (C/Z, m_3) \cdot (T/Z, m_4)$$

Durch ersetzen der beliebigen Basen Y und Z durch jeweils die Base A und das Einsetzen der dazugehörigen m_1 bis m_5 Werte aus obiger Substitutionsmatrix, lässt sich nun beispielsweise folgender P1 Wert berechnen:

$$P1 = (A/A, m_1) \cdot (C/A, m_2) \cdot (A/A, m_5) \cdot (C/A, m_3) \cdot (T/A, m_4)$$

$$P1 = 1 \cdot 0{,}1 \cdot 1 \cdot 0{,}1 \cdot 0{,}1 = 0{,}001$$

Der $L_{1\text{-}1}$-Wert ergibt sich nun aus der Summe der 16 P-Werte. Die allgemeine Berechnungsformel für die Ermittlung eines jeden $L_{i\text{-}j}$-Wertes lautet dabei:

$$L_{i-j} = \sum_{Y = A,C,G,T} P(x_{s1}/Y, m_1) \cdot P(x_{s2}/Y, m_2)$$

$$\cdot \left[\sum_{Z = A,C,G,T} P(Z/Y, m_1) \cdot P(x_{s3}/Z, m_3) \cdot P(x_{s4}/Z, m_4) \right]$$

Eine vereinfachte Darstellung dieser Formel findet sich in Abbildung 13. Auf Grundlage dieser Formel werden nun alle weiteren $L_{1\text{-}1}$ bis $L_{1\text{-}n}$-Wert des ersten Baumes berechnet. Der L_1-Wert dieses Baumes ergibt sich dann wiederum aus dem Produkt aller $L_{1\text{-}1}$ bis $L_{1\text{-}n}$-Wert bzw. der Summe der logarithmierten $L_{1\text{-}1}$ bis $L_{1\text{-}n}$-Wert. Der Grund für das Logarithmieren der $L_{1\text{-}1}$ bis $L_{1\text{-}n}$-Wert liegt darin, die Werte in eine passendere und anschaulichere Form zu bringen. Der Informationsgehalt dieser geht dabei nicht verloren. Somit ermittelt sich der L_1 und $\ln L_1$-Wert nun wie folgt:

$$L_1 = L_{1-1} \cdot L_{1-2} \cdot L_{1-3} \cdot \ ... \ \cdot L_{1-n} = \prod_{i=1}^{n} L_{(i)}$$

$$\ln L_1 = \ln L_{1-1} + \ln L_{1-2} + \ln L_{1-3} + \ ... \ + \ln L_{1-n} = \sum_{i=1}^{n} L_{(i)}$$

Diese ganzen Rechenschritte, werden für die Ermittlung der restlichen L-Werte (im Beispiel wären dies noch der L_2 und L_3-Wert) wiederholt. Am Ende werden alle L bzw. lnL-Werte der einzelnen Bäume miteinander verglichen und derjenige, welcher den höchsten lnL-Wert aufweist, ist jener, welcher die Evolution am wahrscheinlichsten darstellt (der gesuchte ML-Baum). [12][14]

Wie bereits erwähnt, ist die ML Methode die zurzeit am häufigsten angewandte Methode zur Untersuchung evolutionärer Verwandtschaftsbeziehungen. Der Grund dafür ist die Tatsache, dass alle möglichen Wege, welche die Evolution gegangen sein könnte, in Betracht gezogen werden und schlussendlich der wahrscheinlichste daraus ermittelt wird. Den einzigen Nachteil, welche diese Methode besitzt, ist der, dass sie auf Grund ihres enormen Rechenaufwandes eine entsprechend lange Berechnungsdauer besitzt und nur auf Datensätze entsprechender Größe angewendet werden kann.

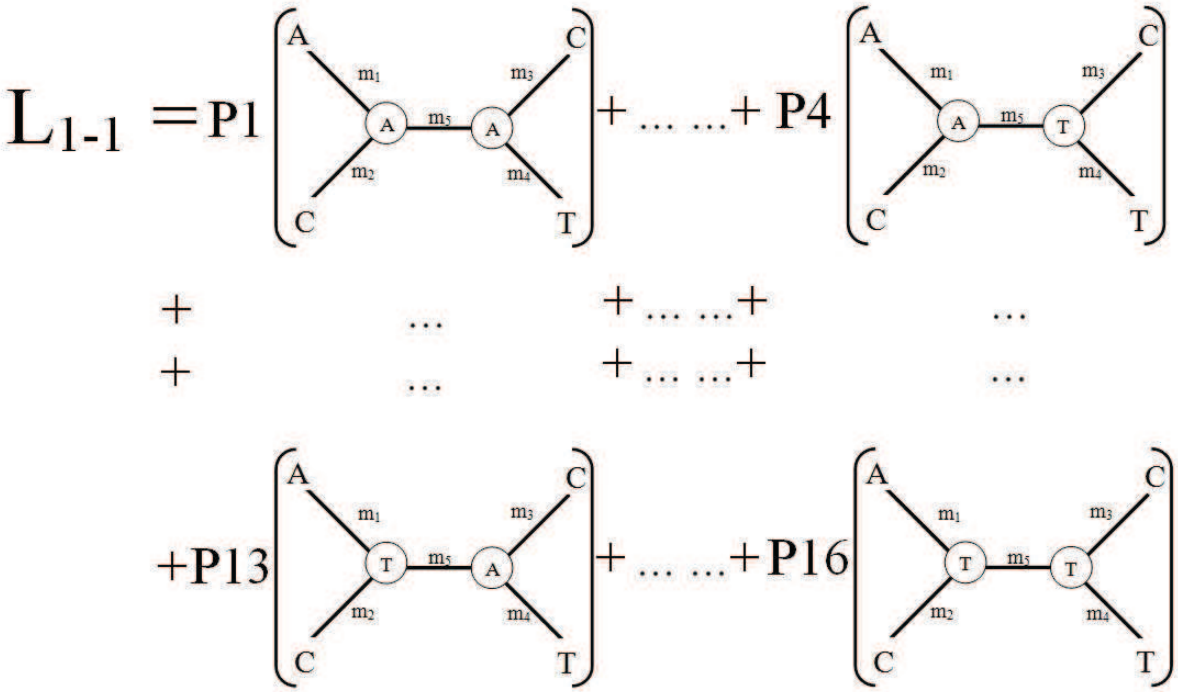


Abbildung 13: A: Schematische Darstellung des Berechnungsablaufes zur Ermittlung des $L_{1\text{-}1}$-Wertes. Die gepunkteten Bereiche repräsentieren die restlichen Bäume, über welche die Mutationen der einzelnen Basen erfolgen können.

4 Die BRCT-Domäne

4.1 Allgemeines

Die Carboxyl-terminale Domäne des Breast Cancer Gens 1 (BRCT-Domäne) ist eine in vielen Spezies weitverbreitete Proteindomäne. Ihr Vorkommen reicht von Bakterien über Pilze und Pflanzen bis hin zum Menschen und einer Vielzahl anderer Eukaryoten. Auch in einigen bisher noch nicht klassifizierten Organismen wurde sie bereits entdeckt. Momentan sind auf der Proteinfamiliendatenbank Pfam mehr als 5400 verschiedene Sequenzen aus 2798 Spezies bekannt. In der Regel weist die Domäne eine Länge zwischen 90 und 100 Aminosäuren auf und sie tritt oft am C-terminalen Ende von Proteinen auf. [15]

Die BRCT-Domäne spielt eine wichtige Rolle bei der Vermittlung von Protein-Protein-Interaktionen oder Protein-DNA-Interaktionen. Dabei nimmt sie vor allem die Funktion eines Gerüstproteins ein und hilft bei der Ausbildung von Multiproteinkomplexen oder unterstützt das Binden der DNA an das Protein. Alle Proteine, welche diese Domäne enthalten, sind entweder direkt oder indirekt an DNA-Transaktionen oder bei der Regulierung dieser beteiligt. Dabei reichen die DNA-Transaktionen von der DNA-Replikation bis hin zu Zellzyklusregulationen, bei welchen DNA-Reparaturaktivitäten auftreten und somit Schäden an der DNA beseitigt werden. [16]

Die Funktion, ob DNA oder Protein bindend, ist je nach Protein unterschiedlich anzutreffen. Oft tritt die Domäne als Tandem-Repeat Sequenz, d.h. als kurz aufeinanderfolgende, durch eine kurze Linkerregion getrennte Sequenz, auf, jedoch kann sie auch als einzelnstehende Domäne vorkommen. So enthält zum Beispiel das DNA-Reparaturprotein XRCC1 zwei voneinander getrennte BRCT-Domäne, von denen jede in der Lage ist eine Proteinbindung zu einem anderen Protein ausbilden. Dies geschieht beispielsweise mit der BRCT-Domäne der DNA-Ligase III oder der BRCT-Domäne der Poly(ADP-Ribose)-Polymerase. Somit besteht die Funktion der BRCT-Domäne in XRCC1 darin andere Proteine zu binden. Die DNA bindende Funktion findet sich beispielsweise in der bakteriellen NAD-abhängigen Ligase, in welcher die Domäne zur Vermittlung von DNA-Bindungen fungiert. Die Bindung zwischen zwei Proteinen muss jedoch nicht zwingend zwischen zwei BRCT-Domänen erfolgen. So bindet z.B. die BRCT-Domäne des Checkpoint Protein

53bp1 an das Brustkrebs verursachende Protein p53, welches keine BRCT-Domäne enthält. [16][17][18]

Allgemein kann man nun die BRCT-Superfamilie in drei Untergruppen unterteilen. Die erste Gruppe besteht aus einem Kern hochkonservierter Domänen, wie sie in Proteinen wie BRCA1, dem *Saccharomyces cerevisiae* Rad9 Protein oder dem p53-Bindeprotein 53bp1 vorkommen. In der zweiten, etwas entfernter verwandten Untergruppe finden sich DNA-bindende Enzyme wie die bereits oben genannte bakterielle NAD-abhängige Ligase oder die Poly(ADP-Ribose)-Polymerase. In der letzten Untergruppe finden sich schließlich die Retinoblastom Tumor-Suppressor-Proteine und ähnliche, welche eine entfernte Verwandtschaft mit der BRCT-Familie aufweisen. [15]

Die ersten beiden Untergruppen werden in der Pfam-Datenbank als BRCT-Familie gekennzeichnet, während die dritte Untergruppe die PTCB-BRCT-Familie bildet, in welcher sich die entfernt verwandten Domänen befinden. Diese beiden Familien bilden zusammen die BRCT-like-Superfamilie. Die Unterteilung der BRCT-Familie kann jedoch auch auf einem anderen Weg erfolgen. Anhand ihres Auftretens im Protein kann man sie einmal als einzeln vorkommende (Single) Domäne oder als doppelt auftretende Tandem Repead Domäne (Double) kennzeichnen. Diese Unterteilung in double Domäne und Single Domäne wird im späteren Verlauf dieser Arbeit noch einmal im evolutionären Kontext genauer erläutert.

4.2 Aufbau und sequenzielles Verhalten

Anhand der bekannten Sequenz und Wechselwirkungen zwischen den Aminosäuren und durch eine Vielzahl an experimentellen Untersuchungen, welche größtenteils mittels Röntgenstrahlenkristallografie (X-Ray) und Kernspinresonanzspektroskopie (NMR) durchgeführt wurden, konnten die Sekundär- und Tertiärstruktur der BRCT-Domäne ermittelt werden. Im Wesentlichen besteht die Domäne aus vier parallelen β-Sheets (β1-β4), welche von drei α-Helices (α1-α3) umgeben sind. Die Topologie bzw. Abfolge der Sekundärstrukturelemente ist: β1α1β2β3α2β4α3 und ist in dem Multi Sequenz Alignement (MSA) in Abbildung 14 zu erkennen.

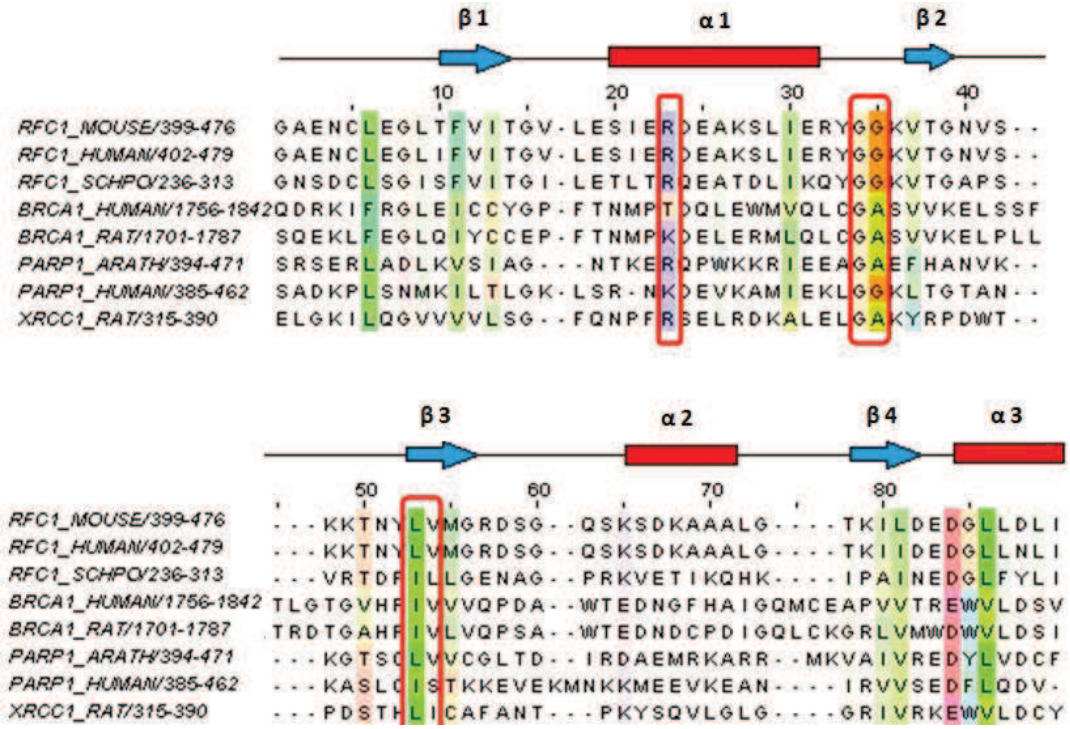


Abbildung 14: MSA für ausgewählte Beispieldomänen und Abfolge der Sekundär Struktur Elemente. Die farbigen Bereiche weisen eine Konservierung von > 30% auf. Je intensiver die Färbung, umso höher ist der Bereich innerhalb der Sequenzen konserviert. Die rot umrandenden Bereiche weisen innerhalb eines MSA von 3810 Sequenzen eine Konservierung von > 50% auf.

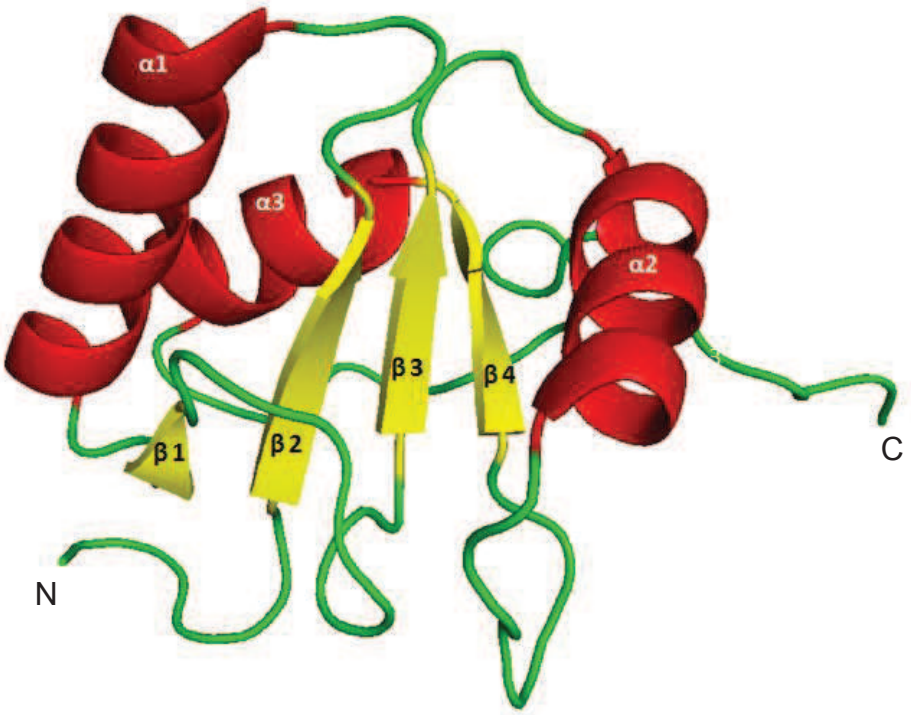


Abbildung 15: 3D-Struktur der BRCT-Domäne aus dem Protein TDT (PDB-Id: 2COE). Gelb dargestellt sind die β-Sheets und rot die α-Helices.

In der Tertiärstruktur ist zu erkennen, dass die vier β-Sheets den Kern der Domäne bilden, welcher von den drei α-Helices umgeben ist. Durch hydrophobe Wechselwirkungen existiert eine große Anzahl von Interaktionen innerhalb der Domäne. So interagiert die α1-Helix mit β1 und β2, während die α2-Helix mit β4 eine Interaktion besitzt. Die konservierten Reste in der α3-Helix wiederum interagieren mit den konservierten Resten in β1, β3 und β4. Diese Interaktionen sorgen für die Stabilität der Domäne. [15][25]

Die BRCT-Familie zeichnet sich vor allem dadurch aus, dass es eine sehr große sequenzielle Vielfalt zwischen den einzelnen Sequenzen der Familie gibt und nur wenige, teils sehr schwach konservierte Bereiche auf sequenzieller Ebene zu erkennen sind. So sind in einem MSA über alle 5443 Sequenzen, welche auf der Pfam Datenbank hinterlegt sind (Pfam ID: PF00533), nur drei Bereiche zu erkennen an welchen eine Konservierung von über 50% auftritt. Diese drei Bereiche bestehen je nur aus ein bis zwei Aminosäuren. Der erste Bereich befindet sich zwischen der α1-Helix und dem β2-Sheet. In diesem Bereich befindet sich ein enger Turn, welcher für die Stabilität der Tertiärstruktur und somit auch für die Funktionalität der Domäne wichtig ist. Dieser kleine hochkonservierte Bereich besteht aus zwei Aminosäuren, wobei die erste Aminosäure in 93% aller Sequenzen ein Glycin ist. Die zweite Aminosäure ist je nach Sequenz mit einer Wahrscheinlichkeit von etwa 90% entweder ebenfalls ein Glycin oder ein Alanin. Experimente aus verschiedenen Studien zeigten, dass eine Substitution des ersten Glycin eine Destabilisierung der dreidimensionalen Struktur zur Folge hat. So zeigte sich z.B., dass eine G435R Mutation in der BRCT-Domäne des RFC Proteins zu einen Rückgang der DNA-Bindeaktivität führt, was wiederum eine Destabilisierung der 3D Struktur zur Folge hat.

Ähnliche, sich negativ auf die DNA bindende Funktion und die Stabilität der Struktur auswirkende Eigenschaften einer Punktmutation von Glycin an gleicher Stelle, wurden bei einer G1788V Mutation der Tandem-Repeat BRCT-Domäne des BRCA1-Proteins und einer G617I Mutation in der BRCT-Domäne der NAD-abhängigen DNA-Ligase beobachtet. [16][17][26]

Der zweite hochkonservierte Bereich befindet sich am Anfang der Sequenz, genauer gesagt innerhalb der α1-Helix. In 70% aller Sequenzen befindet sich an jener Stelle ein Arginin. Die Konservierung dieses Bereiches hängt sowohl mit der DNA bindenden, als auch mit der Protein bindenden Funktion der BRCT-Domäne zusammen. Wie bereits erwähnt ist das Auftreten der Funktion stark vom Organismus abhängig, jedoch scheint die Proteinbindende Funktion weiter verbreitet zu sein. In beiden Fällen spielen jedoch die Seitenketten der Aminosäuren, welche sich am Anfang der α1-Helix befinden, eine wichtige Rolle. Diese Seitenketten sorgen entweder dafür, dass ein gebundener DNA Strang

stabilisiert wird oder gehen Wechselwirkungen mit anderen Proteinen ein, wodurch es möglich ist diese zu binden. Das Arginin scheint an dieser Stelle ein wichtiger Baustein zu sein, da von diesem mehrere Wasserstoffbrückenbindungen (H-Brücken) ausgebildet werden können. [16][17]

Der dritte hochkonservierte Bereich befindet sich in dem β3-Sheet. In diesem Bereich, bestehend aus zwei Aminosäuren, kommt pro Position in etwa 47% aller Sequenzen ein Valin vor. Insgesamt gesehen wird dieser Bereich von vier Aminosäuren dominiert. Valin ist die am häufigsten vorkommende, des Weiteren sind Leucin, Isoleucin und Phenylalanin sehr häufig in diesem Bereich vertreten. Alle vier Aminosäuren sind unpolar und weisen eine hohe Hydrophobizität auf. Diese Eigenschaften sorgen somit für die Stabilität der gesamten Tertiärstruktur. [16][17]

Generell sind die wichtigsten Bereiche, wie schon anhand der Konservierung zu erkennen, das β3-Sheet und der GG-Loop zwischen α1-Helix und β2-Sheet. Diese beiden Stellen sorgen unabhängig von der Funktion, welche die Domäne ausübt, für eine Stabilität der gesamten Struktur. Die funktionell wichtigen Stellen sind zum einen der Anfang der α1-Helix und das β1-Sheet. Bei der α1-Helix dienen, wie schon erwähnt die Seitenketten der jeweiligen Aminosäuren dazu, eine stabile DNA- oder Proteinbindung ausbilden zu können. Aber auch Seitenketten aus der α2-Helix und α3-Helix sind an diesen Bindungen beteiligt, jedoch haben sich deren Funktionen je nach Art der Bindung und Auftreten der Domäne im Laufe der Evolution angepasst. Somit weisen diese Bereiche in einem MSA keine deutliche Konservierung auf.

Das β1-Sheet hat seine Funktion ebenfalls im Laufe der Evolution angepasst. Ursprünglich für die DNA bindende Funktion ausgelegt dient es aber auch bei einigen Domänen dazu, stabile Proteinbindungen ausbilden zu können. Die Rolle welche es dabei einnimmt ist zum Großteil unverändert. Wie die genauen Abläufe der DNA und Proteinbindung erfolgen und welche Bereich der Domäne für diese zuständig sind, soll nun im weiteren Verlauf dieser Arbeit erläutert werden.

4.3 DNA-Bindende Funktion der BRCT-Domäne am Beispiel des Menschlichen Replikationsfaktor C p140

Der Menschliche Replikationsfaktor C (Human Replication Factor C (RFC)) ist einer aus den Untereinheiten p140, p40, p38, p37 und p36 bestehender Proteinkomplex. Als DNA abhängige ATPase ist er in eukaryotischen Zellen für die DNA-Replikation und Reparatur erforderlich. Er wirkt als Aktivator für die DNA-Polymerase, indem er an das 3‘-Ende des Primers bindet und somit die koordinierte Synthese beider Stränge fördert. Es ist bekannt, dass die drei Untereinheiten p40, p37 und p36 für die ATPase Aktivität verantwortlich sind, jedoch ist noch unbekannt welche Untereinheit für die ATP-Hydrolyse zuständig ist. Die Untereinheit p140, welche die größte der fünf Untereinheiten ist, enthält außerdem eine einzelne BRCT-Domäne, welche für die Bindung des Proteins an das 5’-Phosphatende der doppelsträngiegen DNA notwendig ist. [16]

Die BRCT-Domäne von RFC gehört zu einer der deutlichen Unterklassen der BRCT-Familie. Jedoch besitzt sie im Gegensatz zu den meisten anderen BRCT-Domänen eine zusätzliche vorgelagerte α-Helix bei den Resten 375 bis 390, welche durch einen Loop von dem Kern der Domäne getrennt ist. Diese zusätzliche α-Helix (α1‘-Helix) dient speziell in dem RFC Protein dazu, eine zusätzliche Stabilität der gebundenen DNA zu gewährleisten. Experimentelle Untersuchungen aus Studien zeigten, dass das Rückgrat der RFC BRCT-Domäne sowohl im freien als auch im DNA-gebundenen Zustand eine Breite von 1,3Å aufweist, was darauf hindeutet, dass der Kern der BRCT-Domäne während der DNA-Bindung keiner wesentlichen strukturellen Veränderung unterliegt.

In diesen Studien fand man auch heraus, dass die DNA über ihr 5`-Phosphatende an das β1-Sheet der BRCT-Domäne bindet. Dies wurde mit Hilfe eines phosphoreszierten Peptids herausgefunden. Die Kristallstruktur des Komplexes der N-terminalen BRCT-Domäne aus BRCA1 mit einem phosphoreszierten Peptid zeigt, dass die Phosphatgruppe des Peptids an drei Reste des β1-Sheets durch Ausbildung von H-Brücken bindet. Überlagerungen der Struktur dieser N-terminalen BRCT-Domäne mit der Struktur der BRCT-Domäne von RFC zeigen eine deutliche Ähnlichkeit zwischen den Bindungsstellen für das phosphoreszierte Peptid.

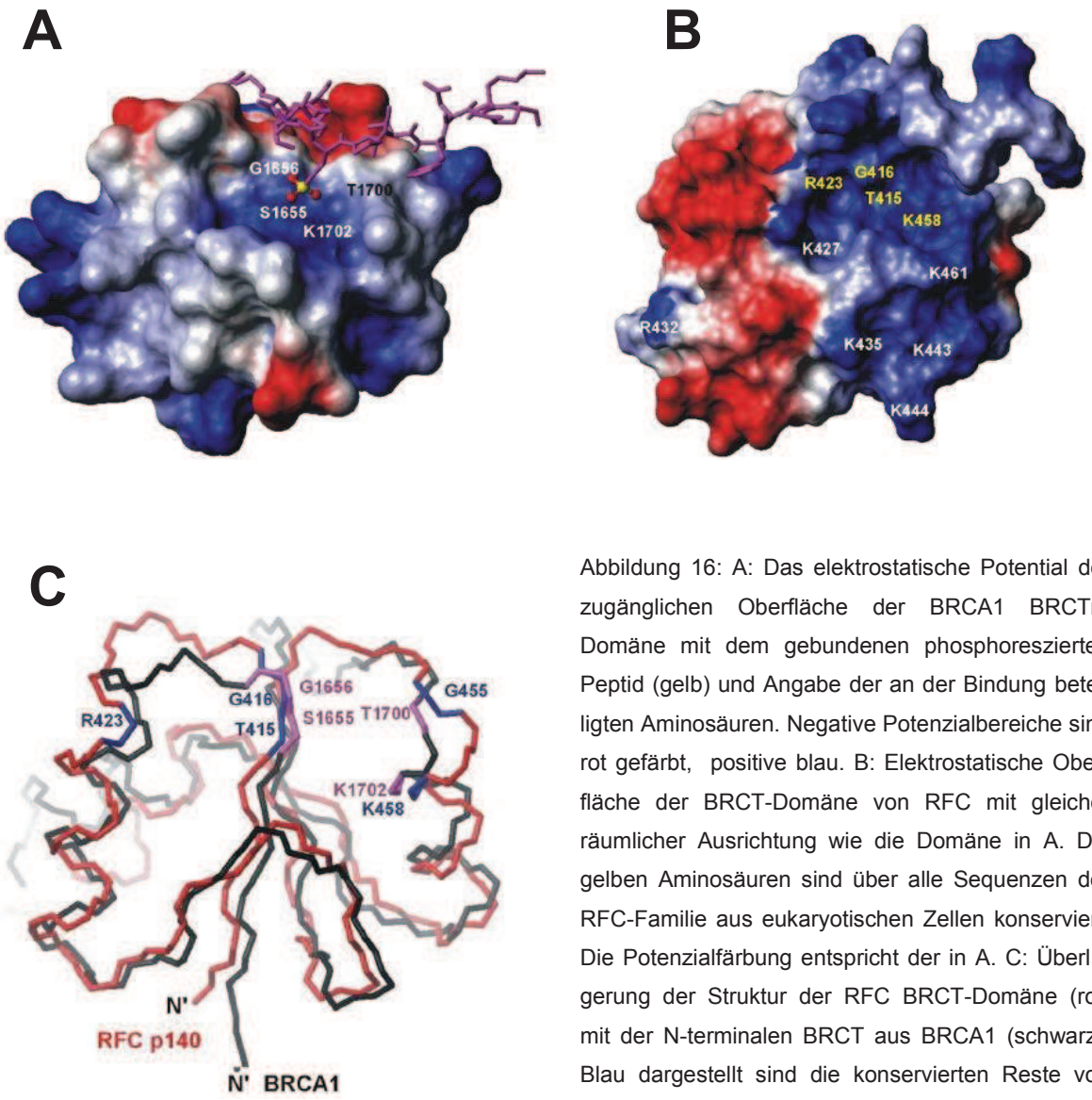


Abbildung 16: A: Das elektrostatische Potential der zugänglichen Oberfläche der BRCA1 BRCTn-Domäne mit dem gebundenen phosphoreszierten Peptid (gelb) und Angabe der an der Bindung beteiligten Aminosäuren. Negative Potenzialbereiche sind rot gefärbt, positive blau. B: Elektrostatische Oberfläche der BRCT-Domäne von RFC mit gleicher räumlicher Ausrichtung wie die Domäne in A. Die gelben Aminosäuren sind über alle Sequenzen der RFC-Familie aus eukaryotischen Zellen konserviert. Die Potenzialfärbung entspricht der in A. C: Überlagerung der Struktur der RFC BRCT-Domäne (rot) mit der N-terminalen BRCT aus BRCA1 (schwarz). Blau dargestellt sind die konservierten Reste von p140. Lila sind die Phosphaterkennenden/-bindenden Reste in BRCA1. [16]

Auch andere BRCT-Domänen weisen eine ähnliche Bindungstasche auf, wie z.B. die N-terminalen BRCT-Domänen aus BRCA1 und MDC1. Auch in diesen Domänen bindet die Phosphatgruppe des Peptides durch Ausbilden von H-Brücken an der gleichen Stelle. Jedoch sind die drei Aminosäurereste, welche die Bindung ausbilden, unterschiedlich. Im Fall von BRCA1 besteht das Trio aus Serin, Glycin und Lysin während in MDC1 an dieser Stelle die Bindung von dem Trio Threonin, Glycin und Lysin, ausgebildet wird. Dennoch weisen all diese Aminosäuren an diesen Positionen einen sehr ähnlichen, leicht negativen Hydrophobizitätswert auf, welcher für die Bindung des Phosphates wichtig ist. Diese spezifizierte Bindung eines phosphoreszierten Peptid bzw. von 5‘-phosphorylierter

doppelsträngiger DNA weist einen konservierten Bereich innerhalb der BRCT-Familie aus, welcher sich am Ende des β1-Sheet befindet. Vor allem die Reste an der ersten Stelle des Trios (im Beispiel Thr-415 bei RFC, Ser-1655 bei BRCA1 und Thr-1898 bei MDC1) scheinen wichtig für das Ausbilden der H-Brücken zu sein. [16][26]

Weitere wichtige Bereiche, welche für die Stabilität der gebundenen DNA sorgen finden sich an Anfang der α1-Helix und der α2-Helix. Die Seitenketten der sich jeweils dort befindlichen Aminosäuren sorgen dafür, dass das gebundene 5'-phosphorylierte DNA Ende bzw. das phosphoreszierte Peptid stabil bindet und keinen größeren Schwankungen unterliegt. In der α1-Helix ist vor allem Arginin ein wichtiger Baustein für diese Stabilität, dementsprechend weist es auch eine relativ hohe Konservierung an dieser Position auf. In der α2-Helix ist die Konservierung der an dieser Stelle befindlichen Aminosäuren weniger ausgeprägt als in der α1-Helix, jedoch weisen sie die gleichen Eigenschaften auf. So besitzen auch diese einen stark negativen Hydrophobizitätswert, ähnlich dem von Arginin, welcher für die Stabilisierung des Phosphors sorgt. In der BRCT-Domäne von RFC übernimmt ein Lysin diese Funktion, während es in BRCA1 ein Threonin ist. In RFC dient außerdem die zusätzliche, vorgelagerte α1'-Helix zu Stabilisierung der gebunden DNA. Dies geschieht dadurch, dass sich der DNA Strang um die α1'-Helix herum windet und somit von den Seitenketten dieser eine zusätzliche Stabilität erhält. Eine Darstellung der Stabilisierung eines phosphoreszierten Peptides in RFC, sowie die Bindung eines DNA Stranges durch die RFC BRCT-Domäne findet sich Abbildung 17. [16]

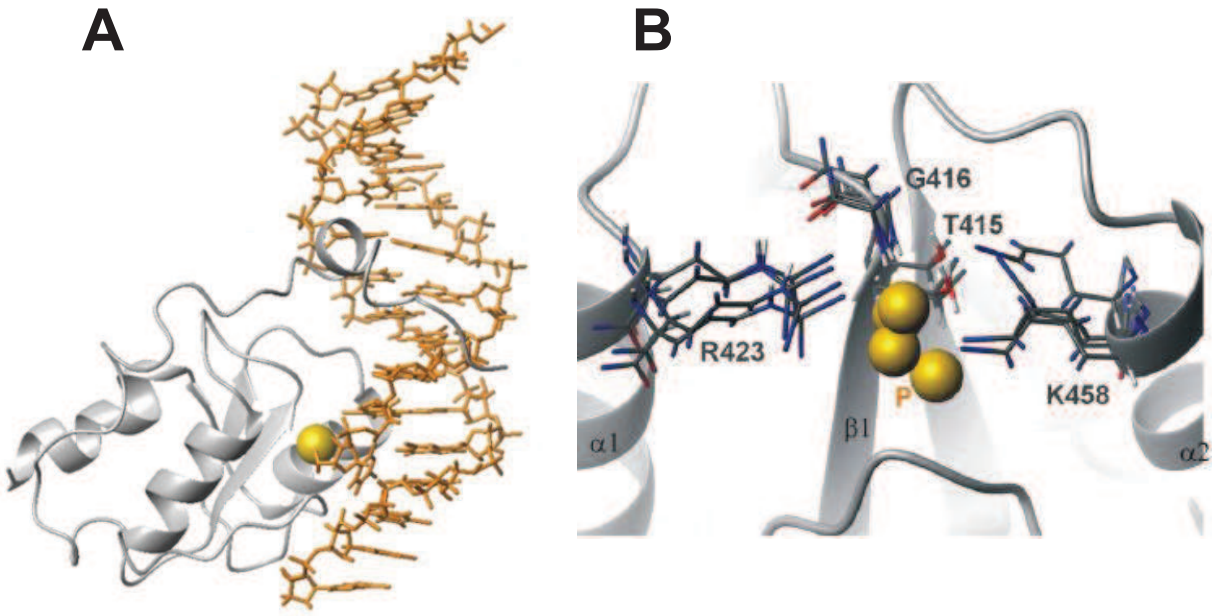


Abbildung 17: A: Struktur der BRCT-Domäne von RFC im DNA gebundenen Zustand. Das Gebundene 5'-Phosphatende ist als gelbe Kogel dargestellt. B: Darstellung der Seitenketten jener Aminosäuren, welche für die Stabilität des gebundenen Phosphates sorgen. [16]

4.4 Protein bindende Funktion durch Dimerisierung zweier BRCT-Domäne am Beispiel der Bindung von XRCC1 mit der DNA Ligase III

XRCC1 (= x-ray cross complementation group 1) ist ein indirekt an DNA-Reparatur- und Reaktionswegen beteiligtes, nicht enzymatisch aktives Protein, welches zwei BRCT-Domänen und eine NTD-Domäne enthält. Es ist ein Gerüstprotein dessen Hauptaufgabe es ist, mit anderen Proteinen Multiproteinkomplexe zu bilden. Dies geschieht sowohl durch die beiden BRCT-Domänen, als auch durch die NTD-Domäne. So bindet beispielsweise an der N-terminalen BRCT-Domäne die BRCT-Domäne der DNA-Ligase III, während die C-terminale BRCT-Domäne eher mit der Poly(ADP-Ribose)Polymerase eine Bindung ausbildet. Aber auch die NTD-Domäne ist in der Lage eine Poly(ADP-Ribose)Polymerase an sich zu binden. Dieser Multiproteinkomplex welcher nun von XRCC1 ausgeht, ist in der Lage die Effizienz des DNA-Reparaturprozesses zu erhöhen, da die jeweiligen Proteine allein so nicht direkt miteinander agieren können und dadurch ineffizienter arbeiten. Die Ausbildung der heterodimeren Bereiche, welche sich dabei zwischen den BRCT-Domänen der jeweiligen Proteine ausbilden findet dabei hauptsächlich zwischen den α1-Helices und den α3-Helices der jeweiligen Proteine statt. [18][19]

In in vito Studien wurden nun die Bindungen welche von der C-terminalen XRCC1 BRCT-Domäne (X1-BRCT) und der BRCT-Domäne aus DNA Ligase III (L3-BRCT) ausgehen, genauer untersucht. Dabei untersuchte man zum einen das Bindungsverhalten, welches zwischen zwei gleichen BRCT-Domänen auftritt und jenes zwischen zwei unterschiedlichen BRCT-Domänen. Grund für die Untersuchung der homodimeren Bindungen, welche durch das Binden von X1-BRCT an X1-BRCT bzw. L3-BRCT an L3-BRCT entstehen ist das Verhalten der jeweiligen an der Bindung beteiligten Seitenketten der Aminosäuren. Die Erkenntnisse aus dieser homodimeren Bindung, wurden mit denen aus der heterodimeren Bindung, welche zwischen X1-BRCT und L3-BRCT entsteht, verglichen, um somit Rückschlüsse auf das allgemeine Bindungsverhalten zwischen Zwei BRCT-Domänen ableiten zu können. [18][19]

Betrachtet man zunächst die homodimere Bindungsstelle zwischen X1-BRCT und X1-BRCT, so stellt man fest, dass die Bindung bevorzugt von der α1-Helix ausgeht. So konnten an dieser Proteinbindungsstelle 19 H-Brücken beobachtet werden. Eine der H-Brücken geht beispielsweise von der Seitenkette von Asp539 aus, wobei diese Seitenkette in den beiden X1-BRCT Monomeren in zwei unterschiedlichen Konformationen vorliegt, was zur Folge hat, dass in Monomer A die Seitenkette aufgrund der veränderten räumlichen Ausrichtung keine H-Brücken ausbilden kann, während dies bei der Seitenkette aus

Monomer B möglich ist. Somit kann das Asp539 aus Monomer B mit Arg562 aus Monomer A zwar 4 H-Brücken ausbilden, aber eine Bindung zwischen Asp539 aus Monomer A mit Arg562 aus Monomer B ist nicht möglich. Somit zeigt sich, dass trotz gleicher Ausprägung der Aminosäuren an diesen Positionen und gleicher Möglichkeiten zur Ausbildung von Bindungen, es durch Fluktuationen in der Ausrichtung der Seitenketten zu unterschiedlichem Bindungsverhalten zwischen gleichen BRCT-Domänen kommt. Vereinfacht ausgedrückt bedeutet dies, dass Bindungen die von Monomer A ausgehen, nicht zwingend auch von Monomer B ausgehen müssen und umgekehrt. Weitere H-Brücken, welche an der Bindungsstelle beobachtet werden konnten, bildeten sich zwischen Arg558 aus Monomer A und den Seitenketten von Glu569, Glu570 und Arg558 aus Monomer B. Die restlichen Bindungen werden von einem Netzwerk von unpolaren Aminosäuren gebildet. Alle Aminosäuren zusammen tragen an der Schnittstelle zu einer gesamten, an der Bindung beteiligten Oberfläche von 1205 $Å^2$ bei. Die genauen Interaktionen an dieser Schnittestelle sind in Abbildung 18 dargestellt. [18][19]

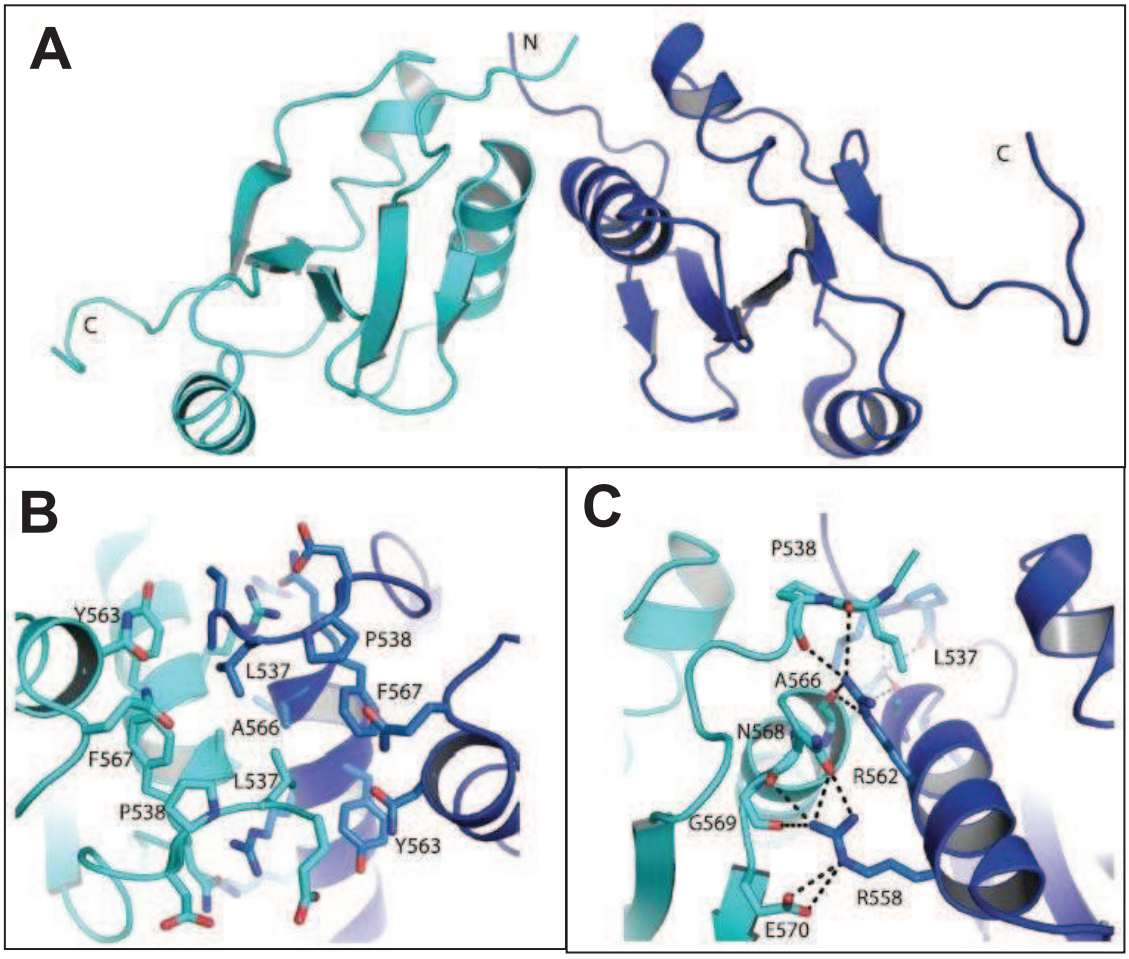


Abbildung 18: Struktur des X1BRCTb Homodimeres. A: Homodimere Bindung beider Monomere. B: Nahaufnahme der hydrophoben Wechselwirkungen. Aminosäure Seitenketten welche an der Bindungsstelle gefunden wurden sind als Stick-Modelle dargestellt. Gezeigt sind nur die hydrophoben Reste. C: Nahaufnahme der H-Brückenbindungen zwischen den beiden Monomeren. [19]

Bei der Untersuchung der homodimeren Bindung zwischen zwei Monomeren aus L3-BRCT wurde festgestellt, dass diese zwei verschieden orientierte Bindungen ausbilden können. Dabei weist die erste räumliche Ausrichtung etwa die gleiche Orientierung auf wie die zwischen den X1-BRCT Monomeren, während bei der zweiten eine um etwa 13° verringerte Orientierung zu erkennen ist. Dies hat zur Folge, dass die beiden Monomere eine geschlossene Struktur bilden und es zu veränderten Interaktionen an der Schnittstelle kommt. An der ersten Bindungsstelle, welche der der X1-BRCT Monomeren ähnlich ist wurden insgesamt 16 H-Brücken nachgewiesen. Das Arg870, welches sich in der L3-BRCT an der gleichen Position befindet wie Arg562 in X1-BRCT, besitzt die Möglichkeit mit sieben potenziellen Partnern H-Brücken zu dem anderen Monomer auszubilden zu können. Diese möglichen Bindungen sind die gleichen, welche man auch in dem konservierten Bindungsnetzwerk von Arg562 in X1-BRCT feststellen konnte. Zusammen mit den restlichen Wechselwirkungen zwischen den unpolaren Aminosäuren beträgt die an der Bindung beteiligte Oberfläche bei dieser strukturellen Ausrichtung 1365 $Å^2$. Auch in dem zweiten Homodimer, welches sich bilden kann, ist das Muster zur Ausbildung der H-Brücken von Arg870 das gleiche, wie in dem ersten Homodimer und in dem von X1-BRCT. Jedoch fallen aufgrund der räumlich veränderten Ausrichtung H-Brücken zwischen Leu847 und Tyr871 weg. Dieser Verlust wird jedoch dadurch kompensiert, dass die Seitenkette von Arg869 mit der Hauptkette von Leu879 und der Seitenkette von Asp878 je eine H-Brücke ausbildet. Durch die veränderte Konformation kommt es natürlich auch zu einer Veränderung der unpolaren Oberfläche an der Schnittstelle. So bildet sich eine zusätzliche hydrophobe Region um Gln881 und trägt zu einer zusätzlichen 50$Å^2$ großen Oberfläche bei. Somit führt diese um 13° veränderte räumliche Orientierung hier zu einer an der Bindung beteiligten Oberfläche von 1304 $Å^2$. [18][19]

Bei der in vitro Untersuchung der Bindungsstelle zwischen X1-BRCT und L3-BRCT stellte sich heraus, dass nicht nur zwischen der α1-Helix und α3-Helix der jeweiligen Domänen sich Bindungen ausbilden, sondern es auch zwischen den Resten des C-Terminus der X1-BRCT zu einer Art Bindung kommt, was dazu führt, dass die im freien Raum befindlichen X1-BRCT und L3-BRCT einen Tetramerkomplex bilden. Dieser Tetramerkomplex besteht im Inneren aus zwei X1-BRCT Domänen, welche über den C-Terminus eine gemeinsame Bindung ausbilden und zwei L3-BRCT Domänen, welche über die jeweiligen α1-Helices und α3-Helices an die X1-BRCTs binden. Die heterodimeren Bindungsstellen zwischen X1-BRCT und L3-BRCT weisen dabei eine sehr große Ähnlichkeit mit denen der homodimeren Bindungsstellen aus X1-BRCT mit X1-BRCT und L3-BRCT mit L3-BRCT auf. Insgesamt befinden sich an der Proteinschnittstelle bis zu 18 potenzielle Partner, zwischen denen es zu einer Ausbildung von H-Brücken kommen kann. Die daran

beteiligten Aminosäuren sind weitestgehend die gleichen, wie sie auch in den homodimeren Bindungsbereichen vorkommen. So bilden Arg558 und Arg 562 aus X1-BRCT je vier H-Brücken zu L3-BRCT aus, während von diesem Arg869 mit der Seitenkette von Glu570 aus X1-BRCT agiert. Die restlichen H-Brücken bilden sich durch Interaktion der Hydroxylgruppe von Tar871 aus L3-BRCT mit den Hauptketten von Pro635 und Leu537 in X1-BRCT aus. Insgesamt beträgt die Größe der an der Bindung beteiligten Oberfläche an diesen heterodimeren Schnittstellen etwa 1210 $Å^2$ und entspricht somit fast genau der Größe, wie sie bei der homodimeren Schnittstelle von X1-BRCT vorkommt. [18][19]

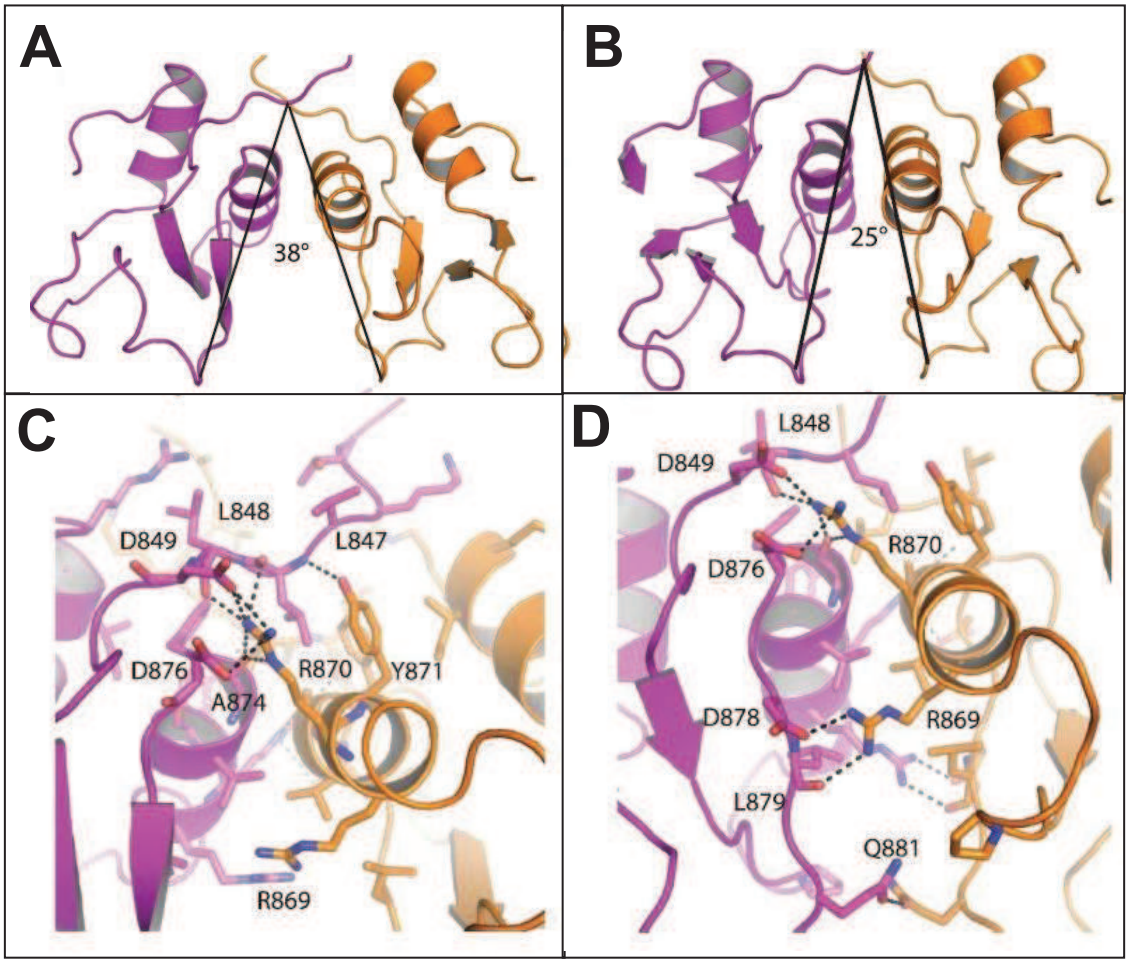


Abbildung 19: Struktur der homodimeren Bindung von L3BRCT. A und B zeigen die beiden verschieden Ausrichtungen an der Bindungsschnittstelle. C: Hydrophyle Wechselwirkungen zwischen den Aminosäuren aus A. D: Hydrophyle Wechselwirkungen aus B. [19]

Bei der in vitro Untersuchung der Bindungsstelle zwischen X1-BRCT und L3-BRCT stellte sich heraus, dass nicht nur zwischen der α1-Helix und α3-Helix der jeweiligen Domänen sich Bindungen ausbilden, sondern es auch zwischen den Resten des C-Terminus der X1-BRCT zu einer Art Bindung kommt, was dazu führt, dass die im freien Raum befindli-

chen X1-BRCT und L3-BRCT einen Tetramerkomplex bilden. Dieser Tetramerkomplex besteht im Inneren aus zwei X1-BRCT Domänen, welche über den C-Terminus eine gemeinsame Bindung ausbilden und zwei L3-BRCT Domänen, welche über die jeweiligen α1-Helices und α3-Helices an die X1-BRCTs binden. Die heterodimeren Bindungsstellen zwischen X1-BRCT und L3-BRCT weisen dabei eine sehr große Ähnlichkeit mit denen der homodimeren Bindungsstellen aus X1-BRCT mit X1-BRCT und L3-BRCT mit L3-BRCT auf. Insgesamt befinden sich an der Proteinschnittstelle bis zu 18 potenzielle Partner, zwischen denen es zu einer Ausbildung von H-Brücken kommen kann. Die daran beteiligten Aminosäuren sind weitestgehend die gleichen, wie sie auch in den homodimeren Bindungsbereichen vorkommen. So bilden Arg558 und Arg 562 aus X1-BRCT je vier H-Brücken zu L3-BRCT aus, während von diesem Arg869 mit der Seitenkette von Glu570 aus X1-BRCT agiert. Die restlichen H-Brücken bilden sich durch Interaktion der Hydroxylgruppe von Tar871 aus L3-BRCT mit den Hauptketten von Pro635 und Leu537 in X1-BRCT aus. Insgesamt beträgt die Größe der an der Bindung beteiligten Oberfläche an diesen heterodimeren Schnittstellen etwa 1210 $Å^2$ und entspricht somit fast genau der Größe, wie sie bei der homodimeren Schnittstelle von X1-BRCT vorkommt. [18][19]

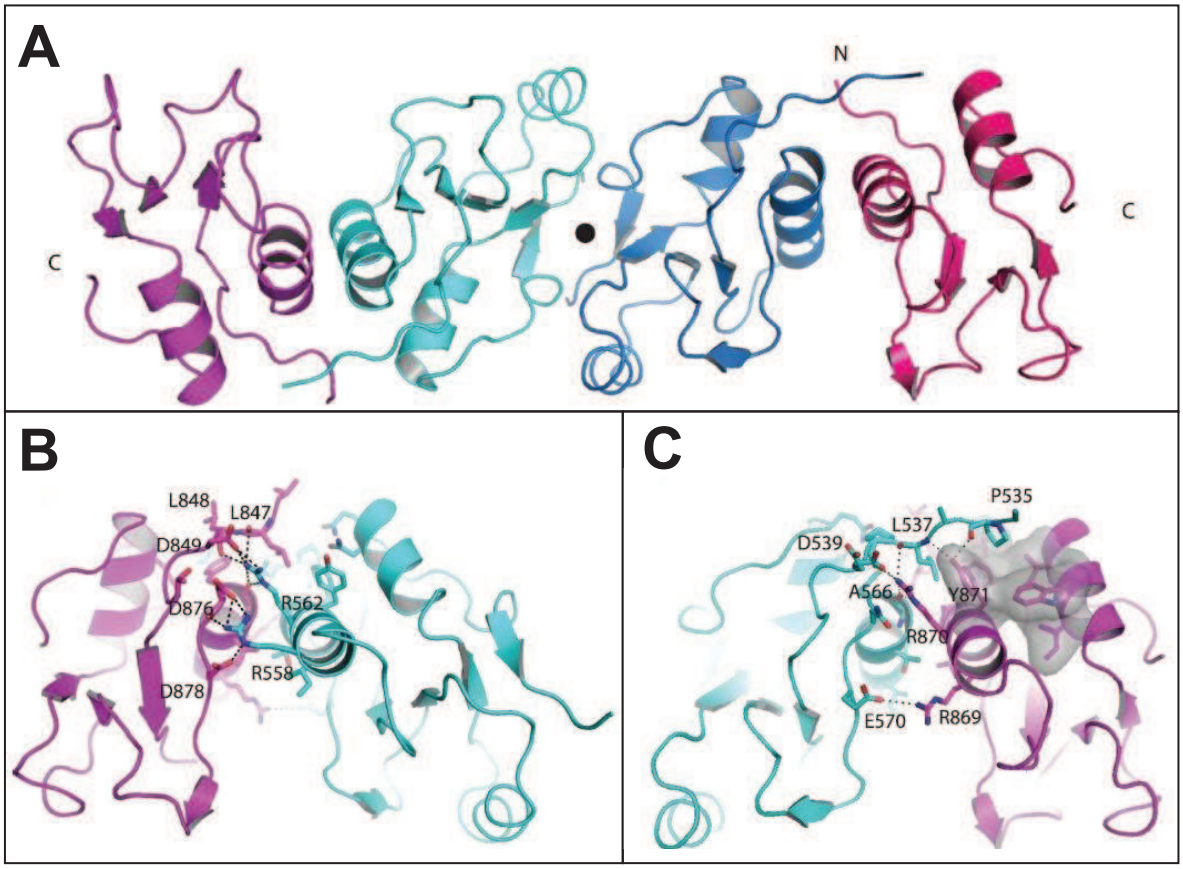


Abbildung 20: A: Tetramerkomplex aus Monomeren von X1BRCT (rot/lila) und L3BRCT (blau). B: Hydrophile Wechselwirkungen an der heterodimeren Schnittstelle zwischen X1BRCTb und L3BRCT. C: Hydrophobe Wechselwirkungen an der homodimeren Schnittstelle zwischen den Monomeren von X1BRCT. Die Wassermoleküle sind als rote Punkte dargestellt [19]

Der Bindungsbereich, welcher sich zwischen den C-Terminus der beiden X1-BRCTs gebildet hat, weist nur sehr wenige H-Brücken auf. Hauptsächlich dominieren an dieser Schnittstelle, welche sich zwischen dem β4-Sheet und einem kleinen helikalen Bereich des Endcoils der jeweiligen Monomere ausgebildet hat, unpolare Wechselwirkungen. In jedem der beiden Monomeren befindet sich ein dicht gepackter Kern aus unpolaren Aminosäuren, bestehend aus Trp588, Phe604, Leu602 und Leu596, welche mit Pro621 und Leu624 aus den jeweiligen anderen Monomeren Van der Waals Wechselwirkungen ausüben. Einige dieser unpolaren Aminosäuren bilden eine Art Tasche, in welcher Wassermoleküle eingeschlossen sind. Diese Tasche ist von beiden Seiten durch unpolare Wechselwirkungen zwischen His622 mit Val624 und Val 627 verschlossen, so dass die Wassermoleküle in dieser „gefangen“ sind. Insgesamt beträgt die an der Bindung beteiligte Oberfläche an dieser Bindungsstelle 1020 $Å^2$. [18][19]

Die Untersuchungen dieser Studie zeigten, dass die Ausbildung von Bindungen zwischen zwei BRCT-Domänen meist zwischen deren α1-Helices erfolgt. Die an diesen Bindungen beteiligten Aminosäuren weisen eine Konservierung innerhalb der BRCT-Familie auf. Besonders Arginin (Arg562 in X1-BRCT bzw. Arg870 in L3-BRCT) scheint eine wichtige Rolle dabei einzunehmen, da von diesem immer die meisten H-Brücken zu der jeweiligen anderen BRCT-Domäne ausgehen und es am meisten Partner zum Ausbilden von Bindungen besitzt. Diese Bindungsbereiche befinden sich meist am Ende der α1-Helix aber auch die Seitenketten von Aminosäuren aus dem ersten Coilbereich der Domäne sind an dieser Bindung beteiligt.

4.5 Protein bindende Funktion durch Interaktion eines Proteins mit der BRCT-Domäne eines andere Proteins anhand des Beispielkomplexes XRCC4/DNA Ligase IV

Der Proteinkomplex, bestehend aus dem Gerüstprotein XRCC4 und der DNA Ligase IV, ist ein wichtiger Bestandteil des sogenannten „nonhomologous end-joining“ (NHEJ) DNA Reparaturmechanismus. Dieser NHEJ Mechanismus ist ein mehrstufiger Prozess, welcher mit der katalytischen Untereinheit der DNA abhängige, Proteinkinas in Wechselwirkung tritt und diese damit bei der Reparatur von DNA Schäden unterstützt. Dabei spielt der XRCC4/DNA Ligase IV Komplex (kurz X4/L4) eine wichtige Rolle bei der Vermittlung der letzten Ligationsschritte. Ein erfolgreiches Binden beider Proteine miteinander ist somit für den gesamten Prozess von großer Bedeutung. Die experimentellen Untersuchun-

gen einer Studie zeigten, wie dieser Proteinkomplex gebildet wird und welche ausschlaggebenden Bereiche der Proteine für die Bindung von Nöten sind. [21]

Diese Untersuchungen zeigten, dass die DNA Ligase IV spiralartig an den langen Helixbereich des XRCC4 Proteins bindet. Dabei sind die N-terminale und C-terminale BRCT-Domäne der DNA Ligase IV um etwa 45° in der Senkrechten zueinander versetzt, wobei die N-terminale Domäne sich unterhalb der C-terminalen in der Ebene positioniert. Die Bereiche, in welchen es zu Ausbildung von Bindungen kommt liegen in XRCC4 zwischen 41-201 und in der DNA Ligase IV zwischen 654-911. Bei der DNA Ligase IV sind diese Bereiche zum einen die α1-Helix und α3-Helix der C-terminalen Domäne und Bereiche in der Linkerregion. Diese Linkerregion ist das markante an der Tandem Repead BRCT-Domäne der DNA Ligase IV, da sie eine außergewöhnliche Länge aufweist. Bei den meisten Tandem Repead BRCT-Domänen schwangt die Länge des Linkers zwischen 5-20 Aminosäuren, jedoch beträgt dieser bei der DNA Ligase IV 60-70 Aminosäuren und weist zwei ausgeprägte α-Helices auf. Aufgrund dieser ungewöhnlichen Länge und der Helix-Loop-Helix Struktur legt sich der Linker klammerförmig um die beiden C-terminalen Helixbereiche von XRCC4 und heftet diese somit aneinander. Diese Art der Bindung im X4/L4 Komplex führt zu einer an der Bindung beteiligten Oberfläche von ca. 4200 $Å^2$, wovon etwa 2000 $Å^2$ auf die Helix-Loop-Helix Klemme zurück zu führen sind. [21]

Bei der genaueren Betrachtung der Interaktionsstelle zwischen der C-terminalen BRCT-Domäne der DNA Ligase IV und XRCC4 ist zu erkennen, dass der größte Teil der Interaktionen in diesem Bereich von der α1-Helix ausgeht und nur einige wenige Seitenketten der α3-Helix beteiligt sind. Dabei weist das Bindungsnetzwerk der α1-Helix eine sehr hohe Ähnlichkeit mit denen auf, welche man auch bei der Dimerisierung zweier BRCT-Domänen miteinander beobachten kann. Dies betrifft sowohl die Dimerisierung zwischen zwei α1-Helices, wie es weiter oben anhand des Beispiels XRCC1-DNA Ligase III beschrieben wurde, als auch jene Dimerisierung, welche zwischen N- und C-terminalen Tandem Repead BRCT-Domänen auftritt, wenn diese über eine kurze Linkerregion verfügen. Bei dem letztgenannten handelt sich jedoch um eine Bindung, welche zwischen der α2-Helix der N-terminalen zu der α1-Helix der C-terminalen ausgeht. Dennoch sind die dort von der α1-Helix ausgehenden Bindungen ähnlich zu jenen im X4/L4 Komplex. Der größte Unterschied zwischen diesen Bindungen ist jedoch die Anzahl der an der Bindung beteiligten Aminosäuren. So sind im X4/L4 Komplex alle Seitenketten, welche von der α1-Helix ausgehen, an der Bindung beteiligt, während bei BRCT-Dimeren je nach räumlicher Ausrichtung meist nur die Seitenketten aus dem Anfangs- bzw. Endbereich der α1-Helix Bindungen ausbilden.

Die Untersuchung der Bindungsaktivität der α1-Helix im X4/L4 Komplex, durch Austausch von Aminosäuren mittels Mutationen zeigte, dass besonders Arg814 eine wichtige Rolle für die Ausbildung einer stabilen Bindung einnimmt. Denn ein unstabiles Binden der C-terminalen BRCT-Domäne zeigt gravierende Auswirkungen auf die Stabilität des gesamten Komplexes. Dieses Arg814 befindet sich, ähnlich wie jene die bei der Bindung des XRCC1/DNA Ligase III komplexes beobachtet wurden, am Ende der α1-Helix. [21] Aber auch andere Aminosäuren finden sich in diesem Bereich, welche sowohl im X4/L4 als auch im XRCCA/DNA Ligase III Komplex vorkommen. Die hohe Ähnlichkeit der Bindungsnetzwerke jener Komplexe weist auf ein deutliches allgemeines Bindungsverhalten der BRCT-Domäne hin, welches in erster Linie von der α1-Helix ausgeht. Diese Art der Proteinbindung ist bei den meisten einzeln auftretenden BRCT-Domänen, wie etwa in XRCC1 oder der DNA Ligase III und auch in einigen wenigen Tandem Repead BRCT Domänen wie in der DNA Ligase IV zu beobachten. Jedoch weisen viele andere Tandem Repead BRCT-Domänen, welche vor allem über eine kurze Linkerregion verfügen, eine andere Art der Proteinbindung auf, welche starke Ähnlichkeit mit der DNA-bindenden Funktion aufweist.

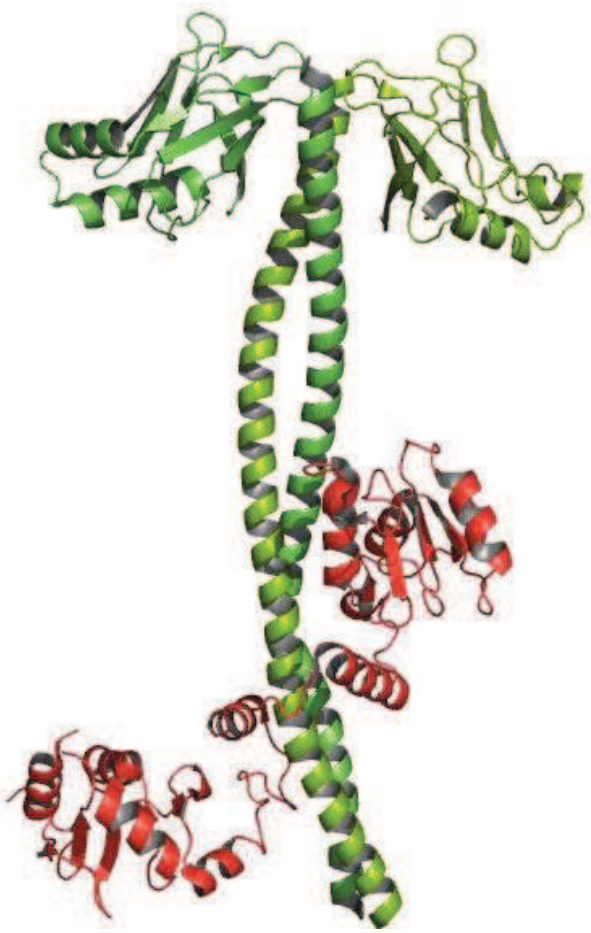


Abbildung 21: Komplex aus XRCC4 und DNA Ligase IV. Die beiden Domänen aus XRCC4 sind grün und die BRCT-Domänen von DNA Ligase IV sind rot gefärbt. Die untere rote Domäne (linke Seite) ist die N-terminale und die obere (rechte) die C-terminale Domäne von DNA Ligase IV. (PDB-Id 3II6)

4.6 Proteinbindende Funktion durch Bindung eines phosphoreszierten Peptides

Die Funktion der BRCT-Domäne lässt sich, wie anfangs bereits erläutert in die DNA bindende und die Proteinbindende Funktion unterteilen. Die Funktion der DNA Bindung wurde bereits ausführlich anhand eines Beispiels beschrieben und auch zwei Möglichkeiten der Proteinbindung, welche auf dem gleichen Prinzip der Ausbildung von Bindungen von der α1-Helix aus beruhen wurden ausführlich erläutert. Es gibt jedoch noch eine dritte Art der Proteinbindung, welche sich die Bindungseigenschaften, welche auch zur Bindung der DNA dienen, zu eigen macht. Gemeint ist damit vor allem die Eigenschaft der Bindung eines phosphoreszierten Peptides in der Bindungstasche, welche von dem β1-Sheet, der α1-Helix und α2-Helix gebildet wird. Diese Art der Bindung kommt vor allem bei Tandem Repead BRCT-Domänen vor, welche durch eine kurze Linkerregion voneinander getrennt sind. Einige solcher BRCT-Domänen finden sich zum Beispiel in dem TopBP1 Protein, aber auch die Tandem Repead BRCT-Domäne von BRCA1 nutzt diese Funktion. Die genau Funktionsweise dieser Bindung soll nun anhand des TopBP1 Proteins näher erläutert werden. [23]

TopBP1 bezeichnet das DNA Topoisomerase II bindende Protein, welches eine wichtige Rolle bei DNA Reparations- und Replikationsmechanismen einnimmt und mit einer Vielzahl von anderen Proteinen bindet bzw. interagiert. Einige von diesen Proteinen sind z.B. die Transkriptionsfaktoren Miz-1 und E2F, das DNA Schadenssensorprotein PARP1 oder, wie der Name schon aussagt, die Topoisomerase II aber auch an Proteinkomplexe wie dem Rad9-Hus1-Rad1 Komplex oder dem BACH/FANCJ Komplex bindet TopBP1. All diese Bindungen gehen dabei von den BRCT-Domänen von TopBP1 aus. Insgesamt befinden sich neun BRCT Domänen in dem Protein, welche in unterschiedlichen Konstellationen auftreten und unterschiedliche Funktionen aufweisen. Bekannt ist bisher, dass alle diese neun BRCT-Domänen zum Binden von Proteinen dienen wobei einige Domänen als Tandem Repead Sequenz vorkommen und einige eher als einzelnstehende. Die genaue Abfolge der BRCT-Domänen in TopBP1 ist in Abbildung 22 schematisch dargestellt. [22][24]

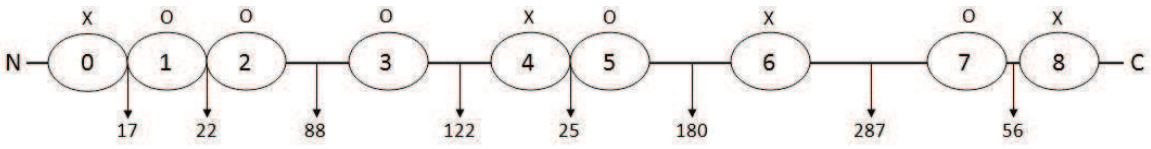


Abbildung 22: Schematische Darstellung der Abfolge der BRCT-Domänen in TopBP1. Die Zahlen unterhalb repräsentieren die Länge der Linker zwischen den Domänen. Das Symbol oberhalb der Domäne steht für die Möglichkeit zur Ausbildung einer Bindung zu einem phosphoreszierten Peptid (O = Bindung möglich, X = Bindung nicht möglich).

Bis auf die dritte und sechste BRCT-Domäne treten die Domänen als Tandem Repead Sequenz auf, wobei das Auftreten der ersten drei Domänen eine besondere Eigenschaft aufweist. TopBP1 ist bisher das einzig bekannte Protein, in welchem eine BRCT-Domäne als sogenannte Triple Sequenz vorhanden ist, d.h. dass drei aufeinander folgende BRCT-Domänen durch je einen kurzen Linker voneinander getrennt sind. Alle bisherigen beobachteten Tandem Repead BRCT-Domänen treten als sogenannte double Sequenzen (zweifach aufeinanderfolgend) auf. Bei genaueren Untersuchungen der einzelnen Domänen stellte man fest, dass die Bindungsstellen für das Binden eines phosphoreszierten Peptides unterschiedlich stark oder gar nicht ausgebildet sind. So weisen, wie in Abbildung 22 zu erkennen ist nur fünf der neun BRCT-Domänen die Eigenschaft dieser Bindungstasche auf, wobei jede Tandem Repead Domäne mindestens eine dieser Taschen besitzt. [24]

Der genaue Prozess der Bindung ist, wie bereits erwähnt, der der DNA Bindung sehr ähnlich. Damit nun ein Protein durch die BRCT-Domäne gebunden werden kann, muss dieses über ein phosphoresziertes Threonin oder Serien verfügen, welche als „Anker“ für das Ausbilden der Bindung dienen. Dieses phosphoreszierte Threonin oder Serien bildet eine Art Äquivalent zu dem phosphoreszierten 5‘-Ende eines DNA Stranges, was zur Folge hat, dass sich im Laufe der Evolution die funktionellen Eigenschaften dieser Bindungsstelle nur minimal verändert und angepasst haben. Somit erfolgt die Hauptbindung des Peptids durch das Binden an das Ende des β1-Sheet und es erfolgt eine Stabilisierung durch die Seitenketten aus der α1-Helix und α2-Helix. Dies zeigen z.B. die Ergebnisse einer Studie, welche die phosphatbindende Eigenschaft der 7/8 BRCT Domäne aus TopBP1 mit dem BACH1 Protein untersuchten. So stellte man fest, dass die Bindung eines phosphoreszierten Seriens aus BACH1 an die 7/8 BRCT-Domäne von TopBP1 durch Wechselwirkungen von Arg1280 aus der α1-Helix, Ser1273 aus dem β1-Sheet und Lys1317 aus der α2-Helix erfolgt. Diese Aminosäuren, welche an dieser Hauptbindung beteiligt sind, weisen innerhalb der BRCT-Familie eine signifikante Konservierung auf. Somit sind, wie auch bei der DNA Bindung, Arginin am Anfang der α1-Helix und Lysin am Ende der α2-Helix ein wichtiger Baustein für die Stabilisierung des Peptides. Ein direkter Vergleich der Bindungstasche der siebten BRCT-Domäne, so wie jene aus der 0/1/2 Triple BRCT-Domäne aus TopBP1 mit der Bindungstasche der DNA bindenden BRCT-Domäne aus RFC ist in Abbildung 23 dargestellt. [22][23]

Anders als bei der Bindung eines DNA Stranges sind bei der Bindung eines Proteins weitaus mehr Aminosäuren beteiligt, da die Proteine eine größere Oberfläche besitzen. Die Untersuchungen der phosphatbindenden Eigenschaft der 7/8 BRCT Domäne aus

TopBP1 mit dem BACH1 Protein zeigten, dass das Binden beider Proteine aneinander zu einer an der Bindung beteiligten Oberfläche von 1208 $Å^2$ führt, was bemerkenswerter Weise in etwa der Fläche entspricht, welche bei der Ausbildung zwischen zwei BRCT-Domänen entsteht. [22]

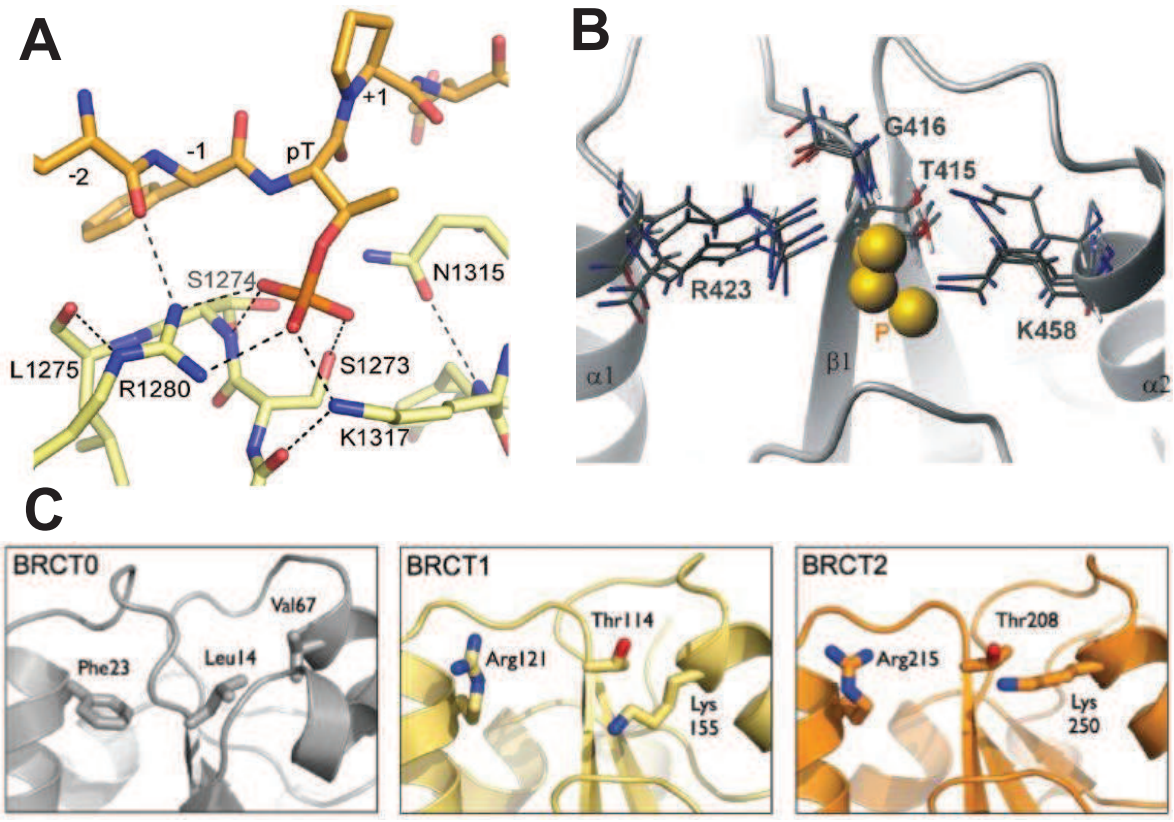


Abbildung 23: Gegenüberstellung der Phosphorpeptid Bindetasche der siebten und der 0/1/2 Triple BRCT-Domäne aus TopBP1 und der BRCT-Domäne aus RFC. A: Bindung eines phosphoreszierten Serins aus BACH1 (Orange) mit der Bindetasche der siebten BRCT-Domäne aus TopBP1 (blasses Gelb). Die Hauptwechselwirkungen zwischen Peptid und Domäne sind als gestrichelte Linien dargestellt. [22] B: Bindung eines phosphoreszierten Peptids (gelbe Kugel) an die Bindungstasche von RFC. [16] C: Seitenketten der Aminosäuren der Bindetasche, aus der nullten, ersten und zweiten BRCT von TopBP1, welche bei der Bindung eines phosphoreszierten Peptids wichtig sind. In der BRCT0 ist keine Bindung möglich, während es bei BRCT1 und BRCT2 möglich ist. [24]

Bei der 7/8 BRCT-Domäne in TopBP1 führt das Binden von BACH1 zu einer Änderung der räumlichen Ausrichtung der achten BRCT-Domäne. Die siebte und achte Domäne sind durch einen verhältnismäßig langen Linker, welcher 56 Aminosäuren lang ist, miteinander verbunden. Diese Linkerregion ist eine einzige große α-Helix, welche als eine Art Rückgrat der double Domänenstruktur gesehen werden kann. Durch diese ist es der achten BRCT-Domäne möglich, bei der Bindung von BACH1 an der siebten BRCT-Domäne eine räumliche Ausrichtung einzunehmen, welche es erlaubt möglichst viele Seitenketten an der Bindung zu BACH1 zu beteiligen. Diese Art der räumlichen Veränderung ist jedoch

bisher nur bei der 7/8 BRCT-Domäne in TopBP1 zu beobachten. So weisen die 5/6 und 1/2 BRCT-Domäne aus TopBP1 keine räumlichen Unterschiede zwischen dem Peptid gebundenen und ungebundenen Zustand auf, genauso wie Proteinpeptid bindende double BRCT-Domänen aus PTIP, BRCA1 und MCPH1.

Eine weitere Besonderheit der Proteinpeptidbindung findet sich in der 0/1/2 Triple BRCT-Domäne von TopBP1. Wie in Studien herausgefunden wurde, besitzen die erste und zweite BRCT-Domäne eine aktive Phosphorbindetasche (siehe Abbildung 22). Bei double BRCT-Domänen besitzt in der Regel nur eine der beiden Domäne eine aktive Bindetasche, während die zweite zur Stabilisierung des gebunden Proteins dient. In der Triple BRCT-Domäne von TopBP1 können jedoch sowohl die erste als auch die zweite Domäne eine stabile Proteinbindung ausbilden. Dies liegt daran, dass die zweite BRCT-Domäne im Vergleich zur ersten um ca. 90° versetzt ist und somit beide Domänen die Stabilisierung eines gebundenen Proteins ermöglichen. Ob beide Domänen jedoch in der Lage sind gleichzeitig zwei Proteine zu binden ist bisher noch unbekannt. Aus Untersuchungen geht hervor, dass sowohl die erste als auch die zweite BRCT-Domäne in der Lage sind, das Gerüstprotein Rad9 zu binden. Durch Untersuchung von Mutationsänderungen in den beiden Domänen stellte man jedoch fest, dass die erste BRCT-Domäne eine festere bzw. effizientere Bindung von Rad9 aufweist als die zweite. Dies lässt die Vermutung zu, dass die zweite Domäne zum Binden eines anderen Proteins dient, jedoch ist bisher noch nicht bekannt, um welches es sich dabei handeln könnte oder dass die zweite Domäne als eine Art Sicherung dient, damit eine sichere Bindung von Rad9 gewährleistet werden kann. [24]

Die Aminosäuren, welche in der ersten und zweiten BRCT-Domäne zum Binden des phosphoreszierten Peptides dienen, sind mit denen aus der siebten BRCT-Domäne und denen aus DNA bindenden BRCT-Domänen, wie zu erwarten ist, sehr ähnlich (siehe Abbildung 23). So treten auch in diesen beiden Domänen Arginin am Anfang der α1-Helix, Lysin am Ende der α2-Helix und Threonin am Ende des β1-Sheets auf. Wie ebenfalls in Abbildung 23 zu erkennen ist, weist die nullte BRCT-Domäne, welche keine aktive Phosphorbindetasche besitzt, keine dieser drei wichtigen Aminosäuren auf. Dies unterstreicht noch einmal die Bedeutung, welche diese drei Aminosäuren bei der erfolgreichen Bindung von DNA und Proteinen besitzen.

4.7 Evolution der BRCT-Domäne

Wie schon des Öfteren in dieser Arbeit angesprochen wurde, haben sich die Funktionen der BRCT-Domäne im Laufe der Evolution an ihre jeweiligen Umgebungsbedingungen angepasst und optimiert. Diese Anpassungen sind über einen langen Zeitraum zurück zu verfolgen und haben ihren Ursprung bei der Entwicklung von einfachen Bakterien zu den komplexeren und höheren Lebensformen der Eukaryoten. In Studien, welche die evolutionäre Entwicklung der Funktion der BRCT-Domäne untersuchten, fand man heraus, dass es drei wichtige Schritte im Laufe der Evolution gab, die dazu führten, dass man die BRCT-Domäne, in diesem Kontext gesehen, in vier große Gruppen unterteilen kann. Diese vier Gruppen werden im weiteren Verlauf dieser Arbeit als Single 1 (s1), Single 2 (s2), Double 1 (d1) und Double 2 (d2) bezeichnet und spiegeln zugleich die funktionellen Eigenschaften, das Auftreten im Protein und den evolutionären Stand wieder. [20]

Am Anfang der evolutionären Entwicklung steht die s1-Gruppe. Die Bezeichnung „Single“ bezieht sich dabei auf das Auftreten der Domäne im Protein. So treten alle BRCT-Domänen, welche in den s1- und s2-Gruppen anzutreffen sind, als markant einzeln vorkommende Domänen im Protein auf. Dies bedeutet nicht, dass nur eine Domäne pro Protein vorkommt. Es können mehrere BRCT-Domänen in einem Protein auftreten jedoch müssen sie durch einen markanten langen Bereich von Aminosäuren voneinander getrennt sein. Ein Beispiel für solch ein Protein ist XRCC1, in welchem die Region zwischen den beiden BRCT-Domänen lang genug ist, um sie nicht als Linker zwischen beiden Domänen zu sehen. Dies würde andernfalls nämlich zu einer Einordnung in die d1- und d2-Gruppen führen. In diesen beiden Gruppen sind alle BRCT-Domänen einzuordnen welche als Tandem Repead Sequenzen vorkommen. Da diese in der Regel immer zweifach auftreten bezeichnet man diese Gruppen als „Double“. Die 0/1/2 Triple BRCT-Domäne aus TopBP1 ist ebenfalls in diese zwei Gruppen einzuordnen, obwohl sie dreifach auftritt. Da TopBP1 das bisher einzig bekannte Protein ist, in welchem die BRCT-Domäne in dieser Triple Konstellation vorkommt und sich die funktionellen und strukturellen Eigenschaften mit denen der double BRCT-Domänen decken wird sie diesen Gruppen zugeordnet.

Wie bereits erwähnt, bilden die BRCT-Domänen der s1-Gruppe den evolutionären Anfang. Alle BRCT-Domäne, welche in diese Gruppe einzuordnen sind, dienen im Organismus dazu DNA zu binden und zu stabilisieren und treten jeweils alleinstehend im Protein auf. Die meisten der Proteine und BRCT-Domänen aus dieser Gruppe sind in Bakterien anzutreffen. Der Grund dafür sind die simpleren DNA-Reparations-, Replikations- und Detektionsmechanismen der Bakterien, im Vergleich zu denen der Eukaryoten. In Bakterien ist die Funktion der DNA-Bindung ausreichend, um die Aufgaben innerhalb der Zelle

zu erfüllen. Mit zunehmender Komplexität der Zellen und Organismen mussten ebenfalls die Mechanismen und Proteine komplexere Aufgaben erfüllen und sich funktionell anpassen. Bei der BRCT-Domäne geschah dies dadurch, dass sich die einfache DNA-bindende Funktion in eine Protein bindende Funktion umwandelte, wodurch es nun möglich ist, Multiproteinkomplexe zu bilden, welche die komplexer werdenden Aufgaben effizienter erfüllen können. Einige dieser DNA-bindenden BRCT-Domänen sind jedoch auch in Eukaryoten noch mit dieser Funktion anzutreffen und sind meist Bestandteil größerer Komplexe wie z.B. die p140 Untereinheit des RFC Protein. [16][20]

Der erste große evolutionäre Schritt war nun jener, welcher zu Aufteilung in die s1- und s2-Gruppe führte. Die entscheidende Veränderung, welche sich in diesem Schritt vollzog war die Veränderung von der DNA-Bindung zu Proteinbindung. Die Art der Proteinbindung entspricht dabei dem Eingehen von Wechselwirkungen zwischen den α1-Helices zweier BRCT-Domänen, wie es im Punkt 4.4 näher erläutert wurde. Typische Vertreter der s2-Gruppe sind z.B. die BRCT-Domänen in XRCC1, DNA Ligase III, DNA Polymerase Mu und die sechste BRCT-Domäne aus TopBP1. [20]

Der zweite große Schritt, welcher jedoch nicht nach dem ersten Schritt, sondern in etwa gleichzeitig mit diesem erfolgte, war die Mehrung der BRCT-Domäne im Protein. So entwickelten sich aus den einzeln vorkommenden BRCT-Domänen doppelt vorkommende Tandem Repead Sequenzen. Jedoch veränderte sich in diesem Schritt nicht nur das Auftretenden im Protein, sondern gleichzeitig auch die Funktion von DNA bindend zu Protein bindend. Die Art der Proteinbindung unterscheidet sich jedoch grundlegend von jener in den s2-Gruppen, denn sie beruht auf der Bindung einer phosphoreszierten Aminosäure eines anderen Proteins wie es in Punkt 4.6 erläutert wurde. Somit besitzen die Proteine dieser Gruppe eine direkte Verwandtschaft mit jenen aus der s1-Gruppe, da sie über den gleichen Bindungsmechanismus verfügen und entwickelten sich somit parallel zu denen der s2-Gruppe. Alle BRCT-Domänen welche sich in diesem Schritt bildeten bzw. die so eben genannten Eigenschaften aufweisen sind in der d1-Gruppe einzuordnen. [20]

Im dritten großen evolutionären Schritt erfolgte eine Art Vereinigung der Eigenschaften, welche sich in den zwei vorhergehenden Schritten ausgebildet hatten. Dies beinhaltet zum einen das Auftreten als double Sequenz und zum anderen die Funktion der Proteinbindung über das Ausbilden von Wechselwirkungen von der α1-Helix aus zu einem beliebigen anderen Protein, so wie es im Punkt 4.5 beschrieben ist. Da die BRCT-Domänen dieser Gruppe, der d2-Gruppe, nun die Eigenschaften dieser beiden evolutionären Schritte aufweisen wird davon ausgegangen, dass deren Entwicklung erst nach der der s2- und d1-Gruppe erfolgte. Ein weiteres Indiz, welches für diese Annahme spricht ist, dass bisher

nur wenige BRCT-Domänen in die d2-Gruppe einzuordnen sind und alle diese bisher nur in höheren entwickelten Eukaryoten wie z.B. dem Menschen vorkommen. Dies spricht dafür, dass sich die Domänen dieser Gruppe nicht direkt aus dehnend der s1-Gruppe entwickelt haben. [20]

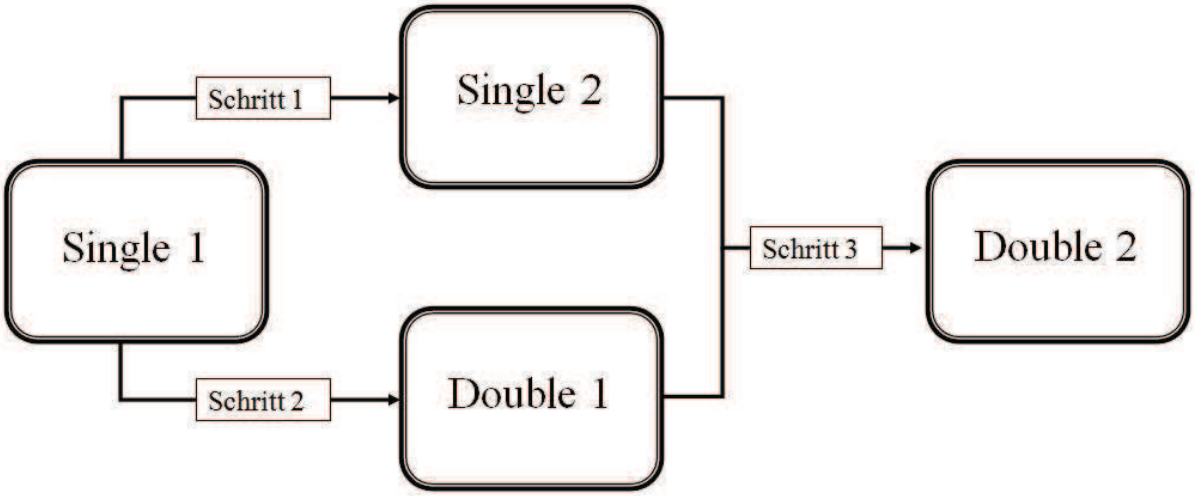


Abbildung 24: Schematische Darstellung des evolutionären Ablaufes der BRCT-Domäne. Eine genaue Beschreibung der Gruppen und Abläufe während der einzelnen Schritte befindet sich Text.

5 Theorie der Energieprofile

Das Prinzip der Protein Energieprofile (EP) beruht auf der einfachen Annahme, dass bei der Betrachtung eines Proteins nicht nur die Abfolge der Sequenz und der Sekundärstrukturelemente, sowie die dreidimensionale Struktur eine Rolle spielen, sondern auch die Energie einer jeden Aminosäure des Proteins. Diese energetischen Zustände der Aminosäuren lassen sich durch ihre jeweiligen physikochemischen Eigenschaften, sowie durch die physikochemischen Interaktionen zu ihrer Umgebung beschreiben.

Wie bereits in Punkt 1.2 erklärt wurde, besitzt jede Aminosäure einen individuellen Rest, welcher der Aminosäure ihre physikochemischen Eigenschaften verleiht. Anhand dieses Restes lassen sich folglich alle Aminosäuren chemisch exakt charakterisieren und klassifizieren. Dies geschieht, wie bereits in Punkt 1.2 dargelegt, oft in Form eines Venn-Diagrammes (siehe Abbildung 2). Aus diesem Venn-Diagramm wiederum kann man nun anhand der Entfernung zwischen zwei Aminosäuren auf die Ähnlichkeiten derer physikochemischen Eigenschaften schließen. Je kürzer die Distanz, umso ähnlicher sind sich diese Eigenschaften. Das gleiche gilt außerdem für die Anziehungs- bzw. Abstoßungskräfte. Somit lassen sich die räumlichen Zustände, welche eine Peptidkette auf dem Weg der Proteinfaltung durchläuft bzw. die anschließend entstandene dreidimensionale Struktur, zum einen durch die physikochemischen Eigenschaften der Aminosäuren und zum anderen durch deren räumliche Anordnung bzw. Distanz zueinander bestimmen. [2][3]

Weiterhin spielen für die Bestimmung der Proteinenergie auch die intermolekularen Wechselwirkungen zwischen den Aminosäuren und dem Lösungsmittel eine große Rolle. Dabei ist die Unterteilung der Aminosäuren anhand ihrer Hydrophobizität eines der wichtigsten Kriterien. Die Hydrophobizität beschreibt das Lösungsverhalten einer Aminosäure (bzw. im Allgemeinen einer Chemikalie) in Wasser oder einem wasserähnlichen Lösungsmittel, wie z.B. Ethanol. Als hydrophob bezeichnet man dabei alle Stoffe, welche sich im Wasser nicht lösen. Hydrophile Stoffe hingegen lösen sich im Wasser, da sie in der Lage sind mit den Wassermolekülen H-Brücken auszubilden. Globuläre Proteine befinden sich sehr oft in Medien bzw. Flüssigkeiten welche einen hohen Wasseranteil aufweisen. Nach der Theorie des hydrophoben Kollapses lässt sich die Interaktion einer Aminosäure mit ihrer Umgebung und dem Protein dahingehend verallgemeinern, indem davon ausgegangen wird, dass sich hydrophobe Aminosäuren häufiger an der Proteinoberfläche befinden, während hydrophile Aminosäuren eher im Proteininneren anzutreffen

sind. Um den energetischen Zustand einer Aminosäure in einem Protein beschreiben zu können, müssen nun folglich alle Aminosäure-Aminosäure- und Aminosäure-Lösungsmittel-Wechselwirkungen erfasst werden. [2][3]

Für die Beschreibung der Aminosäure-Lösungsmittel-Wechselwirkungen wird nun der hydrophobe Charakter der Aminosäure betrachtet. Dabei gilt, dass hydrophobe Aminosäuren dazu neigen in das Innere des Proteins zu wandern, während die hydrophilen Aminosäuren aus dem Inneren mit der Umgebung H-Brücken ausbilden können und sie somit kaum Kräfte zum Molekülinneren ausüben. Dieser Zustand lässt sich als Innen/Außen-Kriterium n(i) definieren. Dabei ist eine Aminosäure (i) als innen ($n_{in,i}$) definiert, wenn gilt:

$$\|c_\alpha - c\| < 5 \vee (c_\alpha - c_\beta)(c_\alpha - c) < 0$$

Dabei entspricht c dem Schwerpunkt im Zentrum des Inneren, um welchen eine Kugel, mit einem Radius r von 5Å gesetzt wurde, um den Innen/Außen-Zustandsraum zu definieren. c_α entspricht der Position des c_α-Atoms und c_β die des c_β-Atoms. Wenn diese Bedingung nicht erfüllt ist, wird die Aminosäure als Außen ($n_{out,i}$) definiert. Unter Verwendung der Boltzmannverteilung (k_B) und der Temperatur (T) lassen sich nun die Lösungsenergien (auch Pseudoenergien genannt) der Aminosäuren berechnen.

$$e_i = -k_B T \ln\left(\frac{n_{in,i}}{n_{out,i}}\right)$$

Da k_B und T jedoch als Konstante angenommen werden, entfallen diese aus der Formel.

$$e_i^* = -\ln\left(\frac{n_{in,i}}{n_{out,i}}\right)$$

Die Energie der paarweisen Interaktion der Aminosäure i zu den anderen Aminosäuren, entspricht der Umgebung (Env) von i und der Umgebungszusammensetzung im Inneren der Struktur. Dies bedeutet, dass die Tendenz der beobachteten Umgebungszusammensetzung (P) mit der Interaktionsenergie von i korreliert. P kann nun durch die abgeleitete Verteilung der Aminosäuresequenzen wie folgt angenähert werden:

$$P_{k\in Env} = \prod_{k\in Env} p_k = \prod_{k\in Env}\left(\frac{n_{in,k}}{n_{out,k}}\right)$$

$$\ln P_{k\in Env} = \sum_{k\in Env} \ln\left(\frac{n_{in,k}}{n_{out,k}}\right)$$

Dabei wird die Umgebung (Env) durch folgende Kontaktfunktion definiert:

$$g(i,j) = \begin{cases} 1, & \|c_\alpha - c\| \leq 8\text{Å} \\ 0, & \text{else} \end{cases}$$

Die 8Å-Umgebung definiert dabei die lokalen und globalen energetischen Einflüsse, welche auf die betrachtete Aminosäure einwirken. Die lokalen Einflüsse sind dabei diejenigen, welche direkt von den benachbarten Aminosäuren ausgehen, während die globalen Einfluss jene sind, welche durch Aminosäuren verursacht werden die sich nicht in unmittelbarer sequenzieller Nähe befinden. Eine Verdeutlichung dieses Sachverhaltes findet sich in Abbildung 25.

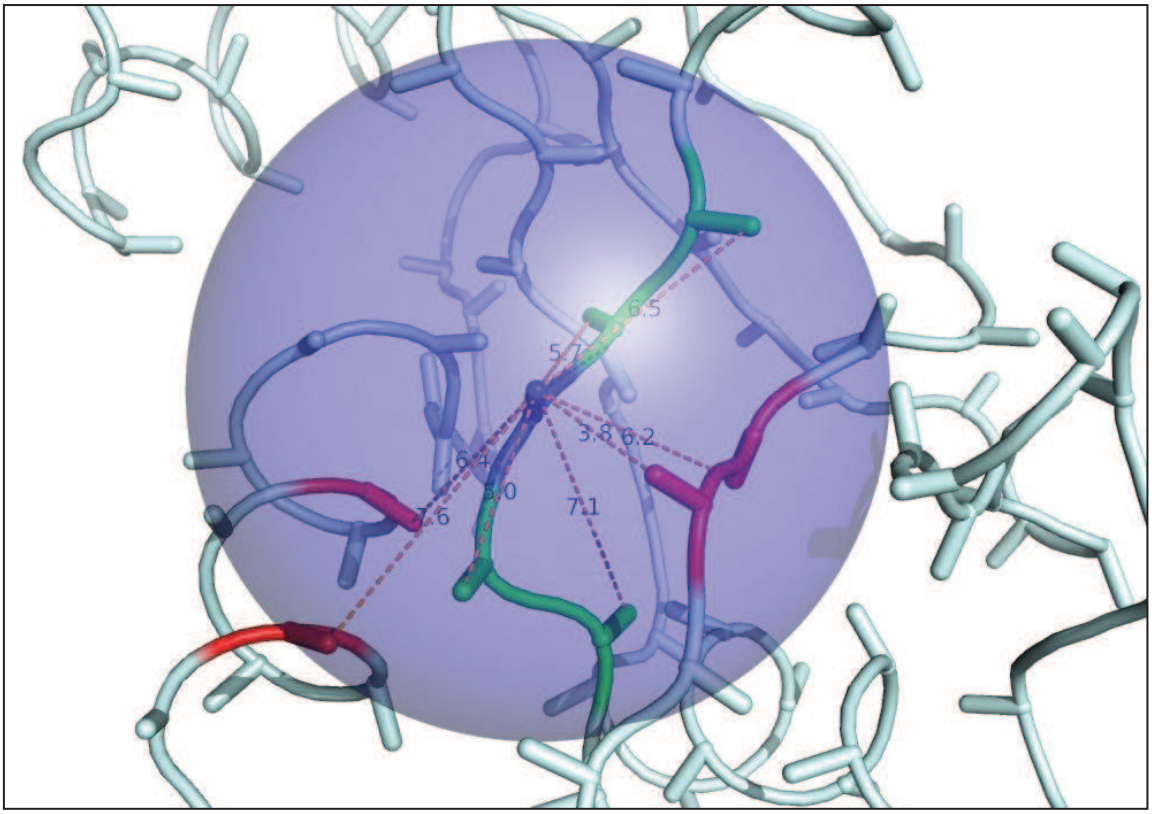


Abbildung 25: Die 8Å-Umgebung um His114 des menschlichen Angiogenius. Die Energie des His114 (blau) wird durch lokale sequenznahe Aminosäuren (grün) und durch globale sequenzferne Aminosäuren (rot) innerhalb der 8Å-Umgebung definiert.

Die Energie der Umgebung (E_{Env}) ergibt sich gemäß dem Boltzmann Prinzip nun wie folgt:

$$E_{Env} = -\ln P_{k \in Env}$$

Somit kann nun die gesamte Energie der Aminosäure i wie folgt ermittelt werden.

$$E_i = -|Env| \ln\left(\frac{n_{in,i}}{n_{out,i}}\right) - \sum_{k \in Env} \ln\left(\frac{n_{in,k}}{n_{out,k}}\right)$$

So ergibt sich schlussendlich folgende Funktion für die Bestimmung der gesamten Energie einer Aminosäure i in einer beliebigen Proteinstruktur (S).

$$E_i^* = \sum_{j \in S \setminus i} \left[g(i,j)(e_i^* + e_j^*) \right]$$

Dieser nun ermittelte Energiewert ist für jede Aminosäure des Proteins einmalig, da zwei Aminosäuren nie durch die gleiche physikochemische Umgebung und Eigenschaften definiert werden. Das entstandene Energieprofil des Proteins ist somit eine zweidimensionale Darstellung des Informationsgehaltes der dreidimensionalen Struktur. Durch Verwendung von EPs erschließen sich nun neue Möglichkeiten der Untersuchung funktioneller und struktureller Eigenschaften von Proteinen. Eine Verwendungsmöglichkeit ist das Alignen von EPs auf Grundlage des Needleman-Wunsch Algorithmus. Ähnlich wie bei dem Alignen zweier Sequenzen werden durch einfügen von Lücken zwei oder mehrere Energieprofile anhand der Ähnlichkeit ihrer energetischen Stellen zueinander angeglichen. Dieses entstandene Multienergieprofilalignment (MEPAL) kann nun wie ein MSA als Ausgangspunkt für weitere Untersuchungen genutzt werden. Ein wichtiger Faktor für weitere Untersuchungen ist dabei die Bewertung von Energieprofilen anhand des dScore. Der dScore ist ein Maß für die Ähnlichkeit zweier Energieprofile zueinander und beschreibt den Editieraufwand, welcher im MEPAL nötig ist um eine EP in ein zweites zu überführen.

Um nun mit Hilfe von Energieprofilen neue plausible Daten zu generieren, wurde in vorangegangenen Studien die Korrelation von strukturbasierenden und sequenzbasierenden Daten zu den Energieprofilen untersucht. So konnte bewiesen werden, dass ein Energieprofil mit der Abfolge der Sekundärstrukturelemente korreliert. So weisen β-Sheets überwiegend niedrige Energien auf, während Coil Bereiche meist hochenergetisch sind. Die α-Helices weisen kein konstantes Energieniveau auf sondern besitzen eine alternierende Abfolge von hohen und niedrigen Energiewerten. Grund dafür ist unter anderem die Lage der α-Helices. Da diese oft am Rand von Proteinen angelagert sind, weisen einige Aminosäuren eine Ausrichtung zum Proteininneren auf, während andere wiederum zum Lösungsmittel hin ausgerichtet sind. Dies sorgt dafür, dass jene, welche eine Ausrichtung zum Proteininneren hin aufweisen, von dem umgebenden Medium isoliert und somit stabi-

ler und niedrigenergetischer sind, da sie keinen Lösungsmittelwechselwirkungen unterliegen. [2]

Weiterhin konnte die Korrelation zwischen struktureller Ähnlichkeit und energetischer Ähnlichkeit nachgewiesen werden. Dafür wurden von sieben Proteinstrukturen, welche keinerlei Ähnlichkeit in Struktur, Sequenz und Funktion zueinander aufwiesen, mit Hilfe des PDBeFold Service die Protein Datenbank (PDB) nach identischen und ähnlichen Proteinstrukturen durchsucht. Für jeden Treffer wurden die Sequenzidentität und der Struktur Alignment Score (QScore) gespeichert und anschließend der Spearman Korrelationskoeffizient zwischen dem QScore, der Sequenzidentität und dem dScore, der abgefragten Proteine berechnet. Diese Korrelationskoeffizienten zeigten, dass zwischen Energieprofil, Sequenzähnlichkeit und struktureller Ähnlichkeit eine hohe Korrelation herrscht. Was wiederum bedeutet, dass eine Transitivität des Energieprofils eines Proteins zu dessen Aminosäuresequenz und Struktur vorliegt. [27]

Anhand dieser Untersuchungen konnte bewiesen werden, dass auf Energieprofilen basierende Daten als Ausgangspunkt für weitere Untersuchungen von Proteinen genutzt werden können. Ein mögliches Einsatzgebiet, welches sich nun daraus erschließen lässt, ist die Untersuchung phylogenetischer Zusammenhänge von Proteinen. Ob es möglich ist, mittels Energieprofilen phylogenetische Zusammenhänge zu rekonstruieren, wie diese zu bewerten sind und wie Energieprofile auf gängige phylogenetische Methoden wie UPGMA, NJ und ML anwendbar sind, soll nun im weiteren Verlauf dieser Arbeit ausführlich erläutert werden.

6 Konstruktion der zu Untersuchenden Stammbäume

Ausgangspunkt für die Untersuchung der evolutionären Zusammenhänge der BRCT-Domäne mit Energieprofil basierenden Daten sind die im Punkt 4.7 dargelegten Erkenntnisse vorangegangener Studien. Ziel dieser Arbeit war es nun, zunächst diese Erkenntnisse auf Grundlage von Energieprofilen zu rekonstruieren und zu überprüfen wie sich Energieprofile auf gängige phylogenetische Methoden anwenden lassen. Den Ausgangspunkt für diese Untersuchungen liefert dabei der Datensatz, welcher dem Fachartikel [20] beilag. Dieser Datensatz umfasst eine Vielzahl an BRCT-Domänen aus verschiedenen Proteinen und verschiedenen Organismen. Mit Hilfe dieses Datensatzes und den im Fachartikel beschriebenen Methoden wurde zunächst überprüft, ob sich die bisherigen evolutionären Erkenntnisse aus dem Fachartikel rekonstruieren lassen, damit eine weitere Anwendung des Datensatzes auf die Rekonstruktion mittels Energieprofilen möglich ist. Die Ergebnisse dieser Überprüfung zeigten, dass sich die Daten bzw. evolutionären Erkenntnisse mit einer signifikanten Ähnlichkeit rekonstruieren lassen und somit der Versuch der Rekonstruktion mittels Energieprofilen fortgesetzt werden konnte.

Zunächst wurde nun der Datensatz dahingehen angepasst, dass ein Arbeiten mit Energieprofilen möglich ist. Wie in Punkt 5 erklärt wurde, ist für die Generierung eines Energieprofils eine zur Sequenz entsprechende Proteinstruktur nötig, welche entweder aus experimentellen Laboruntersuchungen gewonnen wurde oder auf einer der vielen bioinformatischen Datenbanken (z.B. der PDB) vorliegt. Diese Anpassung des Datensatzes offenbarte das erste Problem, welches sich bei dem Arbeiten mit Energieprofilen auftritt. Nach momentanem Stand der Forschung sind in etwa 83100 Strukturen von Proteinen bekannt (Stand der PDB am 24.Juli.2012 [28]). Dies ist, im Vergleich zu den bisher bekannten bzw. entdeckten Proteinsequenzen eine sehr geringe Anzahl. Für die Anpassung des Datensatzes war es nun erforderlich, dass möglichst viele Strukturen von BRCT-Domänen der verwendeten Sequenzen gefunden werden konnten. Dafür wurde in der PDB nach Proteinstrukturen der im Datensatz verwendeten Proteine gesucht. Es stellte sich dabei heraus, dass bisher fast ausschließlich nur Strukturen von BRCT-Domänen aus dem Menschen (*Homo sapiens*) vorliegen. Des Weiteren konnten von einigen Proteinen keine Strukturen gefunden werden, was letztendlich dazu führte, dass nur 18 Proteinstrukturen für das weitere Arbeiten zu Verfügung standen.

In diesen 18 Proteinstrukturen sind insgesamt 31 Strukturen von BRCT-Domänen zu finden, welche teilweise als Single oder Double vorkommend auftreten. Diese 31 BRCT-Domänen sind anhand ihrer Eigenschaften, Funktion und Auftretens laut dem Fachartikel [20] wie folgt in die vier evolutionäre Gruppen (s1, s2, d1 und d2) einzuordnen.

Gruppe	Protein	PDB-Id
s1	RFC	2K6G
	PARP1	2COK
	RAP1	2L42
s2	DNA lig III	3PC8
	DNA poly lamda	2JW5
	DNA poly mu	2HTF
	FCP1	3EF1
	MCPH1 BRCT1	3PA6
	PES1	2EP8
	REV1	2EBW
	TDT	2COE
	TOPBP1 BRCT6	3PD7
	XRCC1 BRCT1	1CDZ
	XRCC1 BRCT2	3PC6
d1	53bp1	1GZH
	BARD1	2NTE
	BRCA1	1JNX
	MCPH1 BRCT2-3	3T1N
	PITP BRCT5-6	3SQD
	TOPBP1 BRCT0-2	2XNH
	TOPBP1 BRCT7-8	3AL2
d2	DNA lig IV	3II6

Tabelle 1: Auflistung aller verwendeten Proteine in denen sich eine oder mehrere BRCT-Domänen befinden. Der Zusatz BRCTx, bei einigen Proteinen gibt an, um welche BRCT-Domänen im Protein es sich handelt. Die Domänen sind in die vier evolutionären Gruppen, welche im Punkt 4.7 erläutert wurden eingeteilt. In der linken Spalte findet sich die PDB-Id. mit welcher die Proteinstruktur auf der PDB hinterlegt ist.

Bei einigen der Strukturen stellte sich zudem heraus, dass diese teilweise Lücken in ihrer Struktur aufwiesen. In den meisten Fällen handelt es sich dabei um Coil Bereiche, welche bei der Kristallisation im Labor mittels NMR oder X-Ray nicht erfasst werden konnten. Da sich diese Lücken auf die Berechnung des Energiewertes negativ auswirken und somit

ein fehlerhaftes EP entstehen würde, wurden diese mit Hilfe des Struktur Modelling Tools des SWISS-MODEL Servers des Swiss Institute of Bioinfomatics [29][30][31] ergänzt.

Von den nun lückenlosen Strukturen konnten mit Hilfe eines in der Programmiersprache Java geschriebenes Programm, welches von Studenten und Mittarbeitern der Hochschule Mittweida programmiert wurde [32], die entsprechenden Energieprofile berechnet werden. Bei der Berechnung der Energieprofile der Double BRCT-Domänen der d1 und d2 Gruppe wurde vorher jedoch die Linkerregion, welche die Double Domänen miteinander verbindet, entfernt. Durch die Entfernung des Linkers wurden die Double Domänen somit in N-terminale und C-terminale Domäne aufgeteilt, so wie es auch bei den Untersuchungen in dem Fachartikel [20] erfolgte. Der Grund dafür ist, dass sich die Funktionen der einzelnen Domänen je nach Vorkommen der Double Domäne unterscheiden und die Aufteilung in N-terminale und C-Terminale Domäne eine bessere Analyse der jeweiligen Funktion und Eigenschaften ermöglicht. Somit wurden von den 31 BRCT-Domänen die entsprechenden 31 Energieprofile generiert.

Um die phylogenetischen Zusammenhänge untersuchen zu können, wurde das von Joe Felsenstein entwickelte Programmpaket PHYLIP [33] benutzt. Dieses Programmpaket, welches als Open Source Datei zur Verfügung steht, ist eine Zusammenstellung von phylogenetischen Analysemethoden und stellt diverse Tools für die Untersuchungen von evolutionären Verwandtschaftsverhältnissen zur Verfügung. Ausgangspunkt für die Konstruktion eines Stammbaumes ist dabei ein MSA bzw. eine Distanzmatrix, welche mit Hilfe des Programmes aus dem MSA berechnet werden kann. Um mit dem Programm arbeiten zu können, müssen sämtliche Eingabedateien in einem speziell für dieses Programm entwickeltem Dateiformat, dem PHYLIP Format, vorliegen. Da bisher noch keine Programme entwickelt wurden, welche auf Grundlage von Energieprofilen eine direkte phylogenetische Analyse ermöglichen, wurde das Arbeiten mit den Energieprofilen dahingehend angepasst, dass diese auf die PHYLIP Programmen angewandt werden konnten. Somit wurde, mit Hilfe eines in Java geschriebenen Programmes [32], aus den Energieprofilen eines MEPAL eine Distanzmatrix im PHYLIP Format generiert, welche von den Programmen des Programmpaketes verarbeitet werden konnte. Mit dieser Ausgangslage konnte nun die Rekonstruktion des evolutionären Verlaufes der BRCT-Domäne erfolgen.

6.1 Anwendung von Energieprofilen auf phylogenetischen Methoden

Da nun eine Anwendung von Energieprofilen auf die Programme des PHYLIP Paketes möglich war, musste nun untersucht werden, welche phylogenetischen Methoden eine Verarbeitung von Energieprofil basierenden Daten zulassen. So stellte sich heraus, dass lediglich die distanzbezogenen Methoden UPGMA und NJ eine Verarbeitung dieser Daten ermöglichen. Charakterbasierende Methoden wie etwa ML können nach momentanem Stand nicht auf Energieprofil basierende Daten angewandt werden. Der Grund hierfür liegt in der Art der Verarbeitung der eingegebenen Daten. Wie bereits im Punkt 3.4 ausführlich erläutert wurde, bezieht der ML Algorithmus seine Daten direkt aus den Substitutionsmatrizen eines MSA, ohne den Zwischenschritt der Generierung einer Distanzmatrix zu vollziehen. In diesen Substitutionsmatrizen sind Wahrscheinlichkeitswerte angegeben, welche anhand des ausgehenden MSA die Wahrscheinlichkeit des Austausches einer Aminosäure durch eine andere beschreiben. Eine derartige vergleichbare Substitutionsmatrize bzw. ein Verfahren für die Berechnung energetischer Substitutionswerte aus Energieprofilen ist momentan noch nicht vorhanden. Somit konnten die weiteren Untersuchungen nur mittels der UPGMA und NJ durchgeführt werden. Dies hatte jedoch zu Folge, dass eine erschwerte Vergleichsgrundlage zwischen den Ergebnisse dieser Untersuchung mit denen des Fachartikels [20] vorlag, da die Ergebnisse des Fachartikels mittels den charakterbasierenden Methoden ML und Bayseian generiert wurden. Wie bereits im Punkt 3 erläutert, stellen UPGMA und NJ die evolutionären Zusammenhänge ungenauer dar als ML, von daher ist eine Projektion von Ergebnissen zwischen diesen Methoden nur schwer möglich. Dies zeigte sich bei dem Versuch, die Ergebnisse des Artikels [20] mittels UPGMA und NJ zu rekonstruieren. So zeigte sich, dass die evolutionären Erkenntnisse und Schlussfolgerungen aus den ML und Bayseian Stammbäumen nicht auch aus den Stammbäumen der UPGMA und NJ Methode gezogen werden konnten, da diese andere evolutionäre Zusammenhänge darstellen.

Nach der Generierung der UPGMA und NJ Bäume auf Grundlage von Energieprofilen (im weiteren Verlauf dieser Arbeit als UPGMA-EP und NJ-EP Bäume bezeichnet) zeigte sich bei dem Vergleich dieser mit denen der auf den Sequenzdaten generierten UPGMA und NJ Bäumen ein weiteres wichtiges Kriterium, welches bei der Bewertung und Beurteilung der Ergebnisse zu beachten ist. Dieses Kriterium bezieht sich auf den Ausgangspunkt für die Berechnung der Stammbäume. Sowohl bei den EP Bäumen als auch bei den Sequenzbäumen ist der Ausgangspunkt eine Distanzmatrix, jedoch unterscheidet sich die Bedeutung der darin stehenden Distanzwerte grundlegend. Wie in Punkt 3 erklärt wurde,

werden die Distanzmatrizen der UPGMA und NJ Methode aus den Substitutionsmatrizen des MSA berechnet. Die Werte dieser Distanzmatrix repräsentieren somit ein evolutionäres Abstandsmaß, welches die evolutionären Verhältnisse der Sequenzen zueinander wiedergibt. In der Distanzmatrix, welche man aus einem MEPAL generiert, repräsentieren die darin stehenden Werte die dScores der einzelnen Energieprofile zueinander. Anders als die Werte der auf Sequenzen basierenden Distanzmatrizen ist der dScore kein evolutionäres Abstandsmaß, sondern beschreibt einen Aufwand, welcher durch Editieroperationen vorgenommen werden muss, um ein Energieprofil in ein anderes zu überführen. Der dScore ist somit ein Maß, welches die Energieprofile im ganzen beschreibt. Auf das MEPAL bezogen bedeutet dies, dass eine globale Betrachtung dieses durch die Distanzmatrix erfolgt. Bei den auf Sequenzen basierenden Distanzmethoden ist dies jedoch anders, da die Substitutionsmatrizen die Mutationswahrscheinlichkeit einer jeden Position des MSA beschreiben. Somit findet eine lokale Betrachtung des MSA statt, da die Veränderungen jeder Position einer jeden Sequenz betrachtet werden. Da eine globale Betrachtung von Daten meinst ungenauer ist als eine lokale Betrachtung lässt sich nun die Schlussfolgerung aufstellen, dass durch die globale Betrachtung des MEPALs die daraus resultierenden EP Bäume die evolutionären Zusammenhänge ungenauer darstellen als die sequenzbasierenden Bäume. Jedoch darf an dieser Stelle ein wichtiger Faktor nicht vergessen werden. Dies ist die Tatsache, dass Energieprofile einen enorm höheren Gehalt an Informationen besitzen als einfache Sequenzen, da sie zusätzlich die Informationen der dreidimensionalen Struktur beinhalten. Daraus erschließt sich nun eine Vorteil/Nachteil Situation, welche zu dem jetzigen Zeitpunkt keine konkrete Aussage über die Qualität der zu erwartenden Ergebnisse ermöglicht. Auf der einen Seite ermöglichen die sequenzbasierenden Daten eine bessere Verarbeitungsgrundlage der Informationen, aufgrund der lokalen MSA Betrachtung. Auf der anderen Seite bieten die Energieprofile aufgrund ihres höheren Informationsgehaltes den besseren Ausgangspunkt für die Untersuchung von Proteinen.

Ein weiterer Vergleichspunkt, welcher für die Analyse der Ergebnisse herangezogen wurde, ist der Vergleich der EP Bäume mit Stammbäumen, welche auf Grundlage der strukturellen Ähnlichkeit der BRCT-Domänen zueinander generiert wurden. Diese strukturellen Stammbäume wurden, wie auch die energetischen und sequenziellen Stammbäume, ausgehend von einer Distanzmatrix berechnet. Die Werte dieser Distanzmatrix geben den pScore der Strukturen zueinander an. Dieser pScore ist ein Maß für die Ähnlichkeit zweier Strukturen zueinander und nimmt dabei einen Wert zwischen null und eins ein. Je ähnlicher sich zwei Strukturen sind, umso kleiner ist der pScore. Folglich zeigt ein Stamm-

baum, welcher aus einer pScore Distanzmatrix berechnet wird, die verwandtschaftlichen Beziehungen auf Grundlage der strukturellen Ähnlichkeit der Proteine/Domänen zueinander an. Die Berechnung der pScores und der diesbezüglichen Distanzmatrix erfolgte mit Hilfe des FATCAT Algorithmus. Dieser wurde mittels eins abgewandelten Java Programmes des BioJava Programmpaketes [34] auf die PDB Dateien des Datensatz angewandt. Der Grund für den Vergleich der strukturbasierenden Bäume mit den EP Bäumen ist der, dass bei Berechnung der Energieprofile die strukturellen Daten mit einfließen. Somit soll überprüft werden, in wie fern die EP Bäume die strukturellen Ähnlichkeiten der Domänen wiederspiegeln.

Die nun resultierenden ungewurzelten UPGMA und NJ Bäume der drei verschiedenen Grundlagen sind in dem Anlagen Teil 1 und Teil 2 zu finden. Welche Korrelation zwischen diesen Stammbäumen herrscht und wie die daraus resultierenden evolutionären Zusammenhänge anhand bisheriger Erkenntnisse zu beurteilen sind, soll nun ausführlich erläutert werden.

7 Untersuchung der Korrelation zwischen Stammbäumen auf sequenzieller, struktureller und energetischer Grundlage

Um die Zusammenhänge und Ähnlichkeiten der Bäume besser untersuchen zu können, wurde eine Färbung der Äste vorgenommen. Diese Färbung steht in Bezug zu den vier evolutionären Gruppen und dient dazu, die verwandtschaftlichen Beziehungen dieser besser erfassen zu können. So wurden die BRCT-Domänen, welche sich in der s1 Gruppe befinden, hellgrün gefärbt und die der s2 Gruppe dunkelgrün. Die Domänen der d1 Gruppe sind als hellrote und die der d2 Gruppe als dunkelrote Äste gekennzeichnet. Des Weiteren wurden OTUs je nach Bedarf und Stammbaum in sogenannte Cluster zusammengefasst. Als Cluster werden somit im weiteren Verlauf dieser Arbeit immer Gruppen von 2-5 OTUs bezeichnet, welche eine signifikante topologische Verwandtschaft zueinander aufweisen und optisch gut von den restlichen Taxa des Stammbaumes getrennt sind.

Für die Analyse der Korrelation zwischen den Bäumen wurde zunächst die Beschaffenheit der drei UPGMA Bäume näher betrachtet. So ist in dem UPGMA-Sequenz Baum der evolutionäre Verlauf in einer etwas „fließenden“ Form dargestellt. In diesem Baum stellt es sich als schwierig heraus, die Veränderungen an den HTUs nachzuvollziehen, welche zu der Aufteilung in die einzelnen Cluster geführt haben könnten. Besonders bei den Domänen der d1 Gruppe ist dies schwierig. Jedoch gibt es zwei Cluster, welche deutlich aus der Baumtopologie herausragen. Dies ist zum einen der Cluster bestehend aus den drei d1 Domänen um BARD1-N herum und zum anderen der angrenzende Cluster bestehend aus vier Domänen der s2 Gruppe um Dpoly Mu herum. Bei der Untersuchung des UPGMA-Struktur und UPGMA-EP Baumes zeigte sich, dass der Cluster um BARD1-N ebenfalls in diesen beiden Bäumen anzutreffen ist. Auch bei dem vierer Cluster der Dpoly Mu ist eine Konsistenz des Cluster zu erkennen. Jedoch beschränkt sich diese auf den zweier Cluster, welcher aus der BRCT-Domäne der Dpoly Mu und TDT besteht. Diese beiden Cluster sind die einzigen, welche in allen drei UPGMA Bäumen anzutreffen sind und dabei keinen wesentlichen Veränderungen unterliegen. Man kann deshalb annehmen, dass sich bei diesen BRCT-Domänen im Laufe der Evolution eine derartig hohe Konservierung ihrer Funktionen und Eigenschaften entwickelt hat, dass diese sowohl auf sequenzieller, struktureller als auch energetischer Ebene zu erkennen sein müsste. Ob dies der Fall ist wird im späteren Verlauf der Arbeit noch genauer dargelegt.

Generell lassen sich die OTUs des UPGMA-Sequenzbaumes auf Grund der fliesenden Baumtopologie nur schwer in Cluster einteilen. Auch bei dem UPGMA-Struktur Baum ist eine Clusterbildung etwas schwierig, jedoch ist sie einfacher als bei dem UPGMA-Sequenz Baum. Bei dem UPGMA-EP Baum jedoch ist eine derartige Einteilung am einfachsten, da in diesen die OTUs optisch klar erkennbare Gruppen bilden. Somit lassen sich in diesem die evolutionären Beziehungen am besten und übersichtlichsten ablesen. Des Weiteren ist in diesem und dem UPGMA-Struktur Baum eine optische Trennung zwischen den BRCT-Domänen der s2 Gruppen und denen der d1 Gruppe zu erkennen. Die OTUs der s1 und d2 Gruppe weisen in allen drei Bäumen unterschiedliche verwandtschaftliche Beziehungen auf, was darauf zurück zu führen ist, dass nur wenige Ausgangsdaten (Strukturen) für BRCT-Domänen dieser Gruppen gefunden werden konnten. Generell kann über diese drei Bäume ausgesagt werden, dass sie nur eine sehr geringe Ähnlichkeit zueinander aufweisen. Die evolutionären Zusammenhänge, welche jeder Stammbaum dargestellt sind nicht auf die anderen beiden Bäume übertragbar. Die einzige Ausnahme bilden dabei die beiden oben angesprochenen Cluster. Des Weiteren sind in dem UPGMA-EP Baum die evolutionären Zusammenhänge am übersichtlichsten dargestellt.

Als nächstes erfolgte die Untersuchung der drei verschiedenen NJ Bäume. So weist der NJ-Sequenz Baum ähnlich wie der UPGMA-Sequenz Baum eine eher „fließende“ Form auf. Des Weiteren zeigt dieser eine größere Vermischung der Taxa der jeweiligen Gruppen, wodurch keine einheitliche Aussage über die Verwandtschaft der Gruppen zueinander getroffen werden kann. Im Gegensatz zu dem UPGMA-Sequenzbaum lässt dieser jedoch ein besseres Clustering seiner OTUs zu. Jedoch weisen die meisten dieser Custer keine Konsistenz innerhalb der drei NJ Bäume auf, so wie es auch bei den UPGMA Bäumen der Fall ist. Die einzigen zwei Cluster, welche in allen drei Bäumen auftreten sind genau wie in den UPGMA Bäumen der dreier Cluster um BARD1-N herum und der zweier Cluster aus TDT und Dpoly Mu. Das Auftreten dieser beiden Clustern in den drei NJ Bäumen festigt die oben aufgestellte Annahme über die evolutionäre Konservierung der Funktionen und Eigenschaften der Domänen. Der Vergleich des NJ-Struktur Baumes mit dem NJ-EP Baum zeigt ebenfalls Parallelen zu den jeweiligen beiden UPGMA Bäumen. So besitzen diese beiden NJ Bäume eine höhere Ähnlichkeit was die Beschaffenheit und Konsistenz der Cluster betrifft zueinander, als zu dem NJ-Sequenz Baum. Jedoch, im Gegensatz zu den UPGMA Bäumen ist bei dem NJ-Struktur und dem NJ-EP Baum die Auftrennung zwischen den s2 und d1 Gruppen nicht mehr so gut vorhanden. So ist diese Teilung in dem NJ-EP Baum komplett aufgehoben, während dies bei dem NJ-

Strukturbaum noch zum Teil zu erkennen ist. Wie auch bei den UPGMA Bäumen zeigen die drei NJ Bäume drei verschiedene Wege, welche die Evolution gegangen sein könnte. Dabei bietet der NJ-EP Baum genau wie der UPGMA-EP Baum die strukturierteste und übersichtlichste Darstellung.

Betrachtet man nun die bisherigen Ergebnisse, so stellt man fest, dass zwischen den Bäumen der drei verschiedenen Grundlagen große Unterschiede und nur wenig Ähnlichkeit bestehen. So kann zu dem jetzigen Zeitpunkt noch keine genaue Aussage darüber getroffen werden, welche der drei Grundlagen die Evolution am plausibelsten darstellt. Um nun aber eine konkretere Aussage zu ermöglichen, wurden jeweils der UPGMA und NJ Baum der jeweiligen Grundlage miteinander verglichen. Da UPGMA und NJ beide auf ähnlichen Berechnungsgrundlagen für die Generierung der Stammbäume beruhen, kann davon ausgegangen werden, dass bei einer hohen Ähnlichkeit der beiden Bäume einer Grundlage zueinander, eine konstante Abbildung der evolutionären Verwandtschaftsbeziehungen vorliegt.

Vergleicht man nun als erstes die Bäume auf sequenzieller Grundlage miteinander, so zeigt sich auf den ersten Blick, dass scheinbar nur eine geringe Ähnlichkeit zwischen diesen vorhanden ist. Bei einer Einteilung der OTUs in Cluster stellt sich jedoch heraus, dass die Cluster der d1 Gruppe eine hohe Konsistenz ihrer verwandtschaftlichen Beziehungen in beiden Bäumen aufweisen. So ist zum einen, wie bereits öfters erwähnt, der dreier Cluster um BARD1-N in beiden Bäumen vorhanden. Des Weiteren weisen zusätzlich der fünfer Cluster um MCPH1-C, der dreier Cluster um TopBP1 0/2-C (welcher die DNA Ligase IV-C Domäne der d1 Gruppe beinhaltet) und der zweier Cluster aus TopBP1 0/2-N und PTIP 5/6-C in beiden Bäumen nahezu die gleichen Verwandtschaftsbeziehungen auf. Bei den Domänen der s Gruppen ist diese Konsistenz von Clustern jedoch nicht in dieser Intensität vorhanden. Dies sorgt für die Vermischung der Taxa im NJ Baum im Vergleich zu dem UPGMA Baum. Der Grund dafür liegt vermutlich in der großen sequenziellen Vielfalt, welche ein charakteristisches Merkmal der BRCT-Domänen ist.

Vergleicht man als nächstes die Topologie der beiden Strukturbäume miteinander, so zeigt sich, dass diese scheinbar eine hohe Ähnlichkeit zueinander besitzen. So sind auf dem ersten Blick keine großen Veränderungen zwischen diesen zu erkennen. An dieser Stelle soll jedoch angemerkt werden, dass diese optische Ähnlichkeit vermutlich auf die Art der Generierung der beiden Bäume zurück zu führen ist. So repräsentieren die Astlängen nicht den direkten pScore, sondern den logarithmierten pScore. Der Grund dafür ist, dass die pScores der Domänen teilweise sehr große Differenzen zueinander aufwie-

sen und dies dafür sorgte, dass so keine übersichtliche Darstellung in Form eines Stammbaumes möglich ist, da einige Taxa enorm weit von dem Zentrum des Stammbaumes entfernt sind. Um nun eine übersichtliche Darstellung zu gewähren, wurden die pScores logarithmiert und anhand eines Grenzwertes bemessen. Dieser Grenzwert repräsentiert einen Wert, ab welchem eine fast 100%igen Strukturidentität angenommen wird und an welchem sich die pScores orientieren. Alle Strukturen, dessen pScores unterhalb dieses Grenzwertes liegen werden somit automatisch als fast 100%ig identisch angesehen und nehmen diesen Grenzwert an. Nach einigen Versuchen zeigte sich, dass ein Grenzwert von $p = 10^{-50}$ (was annähernd Null ist) für die Übersichtlichkeit der Taxa in dem Stammbaum am besten geeignet ist. Die originalen UPGMA und NJ Stammbäume mit den pScores als Astlänge finden sich zum Vergleich im Anlagenteil 3. Aufgrund dieses Grenzwertes weisen die Taxa in den Strukturbäumen fast alle eine annähernd gleiche Länge zueinander auf, jedoch bleibt die Baumtopologie durch das Logarithmieren des pScores unbeeinflusst. Je kürzer nun die Astlänge eines Taxon ist, umso größer ist in diesem Fall die strukturelle Ähnlichkeit der Domänen zueinander. So kann bei der Betrachtung des BARD1-N Clusters davon ausgegangen werden, dass dessen Domänen eine sehr hohe strukturelle Ähnlichkeit zueinander besitzen. Auch die an diesen Cluster angrenzenden Taxa der Domänen der d1 Gruppe, welche in beiden Bäumen gleichermaßen vorhanden sind, scheinen eine deutliche strukturelle Ähnlichkeit zueinander aufweisen zu können. Bei den restlichen Taxa kann jedoch keine genaue Aussage über den Grad der strukturellen Ähnlichkeit getroffen werden, dennoch ist bei genauer Betrachtung der gesamten Topologie der beiden Bäume festzustellen, dass zwischen diesen eine hohe Ähnlichkeit besteht. Diese Ähnlichkeit ist signifikant höher als jene zwischen den UPGMA-Sequenz und NJ-Sequenz Bäumen. Somit kann geschlussfolgert werden, dass die Funktionen und Eigenschaften der BRCT-Domänen eine höhere Konservierung auf struktureller Ebene besitzen als auf sequenzieller Ebene. Ob auch die Bäume auf energetischer Grundlage eine derartige Konservierung aufweisen, wurde als nächstes genauer untersucht.

Sowohl der UPGMA-EP als auch der NJ-EP Baum weisen im Vergleich zu den Sequenz- und Strukturbäumen ein sehr klares und übersichtliches Clustering ihrer OTUs auf. Die Beschaffenheit dieser Cluster ist dabei in beiden Bäumen fast identisch. Bei genauer Betrachtung zeigt sich deutlich, dass sämtliche Cluster der s2 und d1 Gruppe in beiden Bäumen nahezu unverändert vorhanden sind. Lediglich die Taxa der s1 und d2 Gruppe unterliegen etwas größeren Veränderungen, was die Anordnung zu den restlichen Taxa betrifft. Dies kann jedoch unter anderem auf dem Mangel an Ausgangsdaten/Strukturen

zurückgeführt werden. Die Cluster der d1 Gruppen um BARD1-N herum, welche nachweislich in den Strukturbäumen eine sehr hohe Ähnlichkeit ihrer Strukturen zueinander besitzen, sind ebenfalls in den beiden EP Bäumen mit signifikanter Ähnlichkeit vorhanden. Dies ist ein Beleg dafür, dass in den EP Bäumen unter anderem die strukturelle Ähnlichkeit der Domänen mit abgebildet wird und somit die Evolution auf energetischer Ebene mit der auf struktureller Ebene korreliert. Jedoch scheint dies nur bei Strukturen der Fall zu sein, welche eine sehr deutliche bis nahezu identische Struktur zueinander aufweisen. Strukturen mit entsprechend geringerer Ähnlichkeit sind nur schwach zwischen den Bäumen der beiden Grundlagen konserviert. In wie fern nun die strukturelle Ähnlichkeit von den energetischen Eigenschaften beeinflusst wird und ob sich dies positiv oder negativ auf die Darstellung der evolutionären Beziehungen auswirkt, soll im nächsten Kapitel näher erläutert werden.

Zuvor kann nun aber Zusammengefasst, anhand der Analysen der einzelnen Bäume ausgesagt werden, dass zum einen zwischen den Bäumen auf struktureller und energetischer Ebene (sowohl bei UPGMA als auch NJ) eine signifikant erkennbare Korrelation besteht, was die topologische Ähnlichkeit betrifft. Somit sind in den EP Bäumen zum Teil die strukturellen Veränderungen, welche sich im Laufe der Evolution vollzogen haben, abgebildet. Des Weiteren weisen die UPGMA-EP und NJ-EP Bäume die höchste Ähnlichkeit, im Vergleich zu den anderen beiden Grundlagen, zueinander auf. Somit scheint auf energetischer Ebene eine starke Konservierung der Funktionen und Eigenschaften der Domänen vorzuliegen. Die verwandtschaftlichen Beziehungen dieser Funktionen und Eigenschaften sind zudem auf einer gut übersichtlichen und leicht erfassbaren Art in den Stammbäumen abgebildet, was eine erleichterte Analyse dieser zulässt. Somit kann an dieser Stelle die Behauptung aufgestellt werden, dass auf Energieprofilen basierende phylogenetische Stammbäume, im Vergleich zu bisherigen sequenzbasierenden Stammbäumen, eine bessere und übersichtlichere Darstellung der verwandtschaftlichen Beziehungen ermöglichen, in welchen zusätzlich Merkmale der strukturellen Ähnlichkeit zu finden sind. Zudem scheint es so, dass der enorme Informationsgehalt der Energieprofile dafür sorgt, dass der Nachteil der globalen Betrachtung bei der Generierung der Stammbäume aufgehoben wird. In wie weit nun eine lokale Betrachtung Auswirkungen auf die bisher präsentierten Ergebnisse hätte, kann jedoch nicht gesagt werden, da, wie bereits erwähnt, momentan keine lokale Betrachtung durch eine energetische basierende Substitutionsmatrize möglich ist.

Um nun die soeben aufgestellte Behauptung festigen zu können wurde eine ausführliche Untersuchung der verwandtschaftlichen Beziehung der EP Bäume vorgenommen. Ziel

war es dabei, unter anderem eine Plausibilität für die Darstellung der verwandtschaftlichen Beziehungen zu beweisen und neue Erkenntnisse über die funktionellen Veränderungen, welche sich im Laufe der Evolution innerhalb der BRCT-Domänen auf energetischer Ebene vollzogen haben, zu gewinnen.

8 Untersuchung des Verlaufes der Evolution der BRCT-Domänen auf energetischer und struktureller Ebene

Im Laufe der bisherigen Untersuchungen wurden einige Behauptungen aufgestellt, deren Gültigkeit bzw. Plausibilität nun genau ergründet und bewiesen werden muss. Zum einen war es erforderlich, dass ein Beweis für die Plausibilität der evolutionären Darstellung der EP Bäume erbracht werden sollte. Zudem war es wichtig herauszufinden, in welchem Verhältnis die Erkenntnisse aus diesen Bäumen zu denen der bisherigen evolutionären Erkenntnisse der BRCT-Domäne zu betrachten sind. Des Weiteren war eine Untersuchung der Veränderungen der funktionell wichtigen Stellen, auf welche ausführlich im Punkt 4 eingegangen wurde, von Bedeutung. Dabei sollte zum einen die energetische Konservierung dieser Stellen überprüft werden und in welchem Bezug diese zu strukturellen und sequenziellen Merkmalen und evolutionären Veränderungen stehen. Zudem sollte überprüft werden, ob Veränderungen auf energetischer Ebene dazu geführt haben könnten, dass eine evolutionäre Aufteilung der BRCT-Domänen erfolgte. Dies beinhaltete ebenfalls, ob diese energetischen Veränderungen dabei einen Einfluss auf die strukturelle oder sequenzielle Ebene aufweisen.

Diese Untersuchungen wurden anhand des NJ-EP Baumes durchgeführt. Da NJ die etwas genauere phylogenetische Methode ist als UPGMA, ist die Untersuchungen anhand dieses Baumes die logischere Wahl. Um nun diese genaueren Untersuchungen durchführen zu können, erfolgte zunächst eine Einteilung der Taxa des NJ-EP Baumes in Cluster. Eine genaue Beschreibung dieser Einteilung befindet sich in der nachfolgenden Tabelle und eine größere Abbildung des NJ-EP Baumes ist im Anlagenteil 4 zu finden.

Cluster	Domänen	Evolutionäre Gruppe
C1	BARD1-N	d1
	BRCA1-N	d1
	MCPH1 2/3-N	d1
C2	PTIP 2/3-N	d1
	TopBP1 0/2-M	d1
	TopBP1 7/8-N	d1
	Dlig IV-C	d2

C3	BARD1-C	d1
	PTIP 2/3-C	d1
	TopBP1 0/2-N	d1
	53bp1-C	d1
C4	53bp1-N	d1
	BRCA1-C	d1
	TopBP1 7/8-C	d1
C5	TDT	s2
	Dpoly mu	s2
	REV1	s2
	PES1	s2
	Dpoly lam	s2
C6	MCPH1-s	s2
	TopBP1 6	s2
	TopBP1 0/2-C	d1
	XRCC1-s-N	s2
	XRCC1-s-C	s2
C7	Dlig III	s2
	MCPH1 2/3-C	d1
	FCP1	s2
	Dlig IV-N	d2
	RFC1	s1

Tabelle 2: Einteilung der BRCT-Domänen in Cluster. In der rechten Spaltet steht die entsprechende evolutionäre Gruppe in welcher die Domäne einzuordnen ist.

Die Analyse der Cluster erfolgte anschließend mit Hilfe von MEPALs und dem Struktur Visualisierungsprogramm PyMol [25], welches als Open Source Programm im Internet zur Verfügung steht. Für die Analyse der evolutionären Veränderungen wurde nun zunächst für die OTUs des betrachteten Clusters ein MEPAL generiert. Zusätzlich erfolgte ein Strukturaligning der jeweiligen Strukturen in PyMol. Mit Hilfe eines Javaprogrammes [32] konnten die entsprechenden PDB Dateien im Vorfeld so bearbeitet werden, dass es möglich war, in PyMol eine der Energien entsprechende Färbung der Strukturen vorzunehmen. Mit Hilfe der energetisch eingefärbten Strukturalignments und den dazugehörigen MEPALs war es nun möglich herauszufinden, welche Veränderungen auf welche Ebene dazu geführt haben, dass an den HTUs eine Aufspaltung zwischen den Domänen erfolgte. Zunächst wurden dabei die HTUs innerhalb der Cluster untersucht und anschließend jene HTUs, welche sich zwischen den Clustern befinden. Ein besonderes Augenmerk lag dabei auf den HTUs, an welchen sich die Domänen in zwei evolutionäre Gruppen aufspalten. Veränderungen, welche an solch einer HTU stattfanden und könnten sich nachweis-

lich in dem Großteil aller Domänen einer evolutionären Gruppe finden lassen und somit evtl. als evolutionäre Marker dienen. Diese evolutionären Marker könnten bei der Identifizierung unbekannter BRCT-Domänen behilflich sein, indem sie Hinweise über das Vorkommen, Auftreten und die Funktion der unbekannten Domäne liefern.

Bevor nun die einzelnen Cluster näher betrachtet wurden, fand zunächst eine Betrachtung der MEPALs der einzelnen evolutionären Gruppen statt. Dabei zeigte sich, dass in diesen gut die energetische Konservierung der β-Sheets als niedrigenergetische Bereiche zu erkennen ist. Besonders der Kern des β3-Sheets weist dabei eine sehr deutliche und markante Konservierung auf. Bei der weiteren Betrachtung ist ebenfalls ein hochenergetischer Bereich am Anfang des MEPALs der Double Gruppen gut zu erkennen. Bei diesem Bereich handelt es sich um den Anfang der α1-Helix, welcher auch bei den Domänen der anderen Gruppen eine entsprechende hochenergetische Konservierung an dieser Stelle aufweist. An dieser Stelle stößt man jedoch bei der energetischen Analyse allein durch MEPALs bereits an Grenzen. Durch das Einfügen von Lücken kommt es nämlich sowohl in einem MEPAL, als auch in einem MSA zu einer Verschiebung der Struktur. Dies bedeutet, dass Bereiche, welche in einem MEPAL oder MSA übereinander liegen und evtl. eine hohe Konservierung aufweisen, nicht zwangsläufig auch in einem Strukturalignment übereinander liegen müssen. So kommt es in MEPALs vor, dass die Energien eines Bereiches, welche z.B. eine hochenergetische Konservierung aufweisen, in Wirklichkeit innerhalb der Strukturen nicht an ein und derselben Stelle sondern an verschiedenen Stellen anzutreffen sind. Je mehr Objekte in einem MEPAL vorhanden sind und je größer die Unterschiede in der Länge der Sequenzen sind, umso größer ist die Differenz zwischen den energetischen Bereichen des MEPALs und der wahren Positionen innerhalb der Strukturen. Um mit Hilfe von MEPALs eine bessere Analyse evolutionären Veränderungen zu ermöglichen, wurden diese hauptsächlich für die Betrachtung der einzelnen Cluster und für die zwischen zwei Clustern benutzt. Da für die Untersuchungen der funktionell wichtigen Stellen der Domänen auf ihre energetischen Eigenschaften hin jedoch eine genauere Betrachtung von Nöten ist, wurden diese mit Hilfe des Strukturalignment der energetisch eingefärbten Strukturen durchgeführt.

8.1 Analyse der N-terminalen Double BRCT-Domänen

Der erste Cluster, welcher nun einer genaueren Analyse unterzogen wurde, ist der bereits öfters angesprochene dreier Cluster bestehend aus BRAD1-N, BRCA1-N und MCPH1 2/3-N. Das Vorgehen bei der Analyse eines jeden Cluster war dabei immer, dass zu-

nächst eine Betrachtung des MEPALs erfolgte. Dabei wurde nach Stellen gesucht, in denen deutlich erkennbare Veränderungen, der energetischen Zustände erfolgten. Um zu überprüfen ob diese Veränderungen tatsächlich an ein und derselben Stellen innerhalb der Strukturen vorkamen, wurden diese anschließend mittels des Strukturalignments überprüft. So zeigte sich nach der Überprüfung der energetischen Stellen im Strukturalignment, dass es bei BARD1-N zu einem starken Energieabfall am Ende der α1-Helix gekommen war. Des Weiteren zeigte sich, dass es im oberen Bereich der α1-Helix ebenfalls ein starker Energieabfall bei BRCA1-N zu erkennen ist. An dieser Stelle befindet sich eine der drei wichtigen Aminosäuren, welche Bestandteil der Bindetasche für das Binden eines phosphoreszierten Peptides oder DNA Stranges ist. Dank bisheriger Untersuchungen dieser Bindungstasche und dem Wissen darüber, an welchen Stellen diese drei Aminosäuren innerhalb der Struktur vorkommen (siehe Punkt 4), ist es möglich, eine genaue Untersuchung der Energien dieser Bindungstasche vorzunehmen. Somit zeigte sich, dass die Aminosäurereste dieser Bindungstasche in allen drei Domänen des Clusters eine hochenergetische Konservierung aufweisen. Nur der Bereich in der α1-helix in BRCA1-N weist stark negative Energien auf. Da hohe Proteinenergien auf eine starke negative Hydrophobizität und polare Aminosäuren hinweisen und man weiß, dass diese Eigenschaften wichtig für das Ausbilden von Bindungen zu einen Peptid oder DNA Stranges sind, kann anhand dieser Energiewerte ausgesagt werden, dass die drei Domänen über eine aktive Phosphorbindetasche verfügen. Da jedoch in BRCA1-N eine starke energetische Veränderung in einem Bereich dieser Tasche stattgefunden hat, könnte man die Hypothese aufstellen, dass in dieser Domäne eine Veränderung der Bindungseigenschaften erfolgt.

Eine ähnliche energetische Veränderung innerhalb dieser Bindungstasche ist ebenfalls in dem Cluster 2 zu erkennen. So zeigt sich, dass bei PTIP 5/6-N sowohl an der Stelle der Bindungstasche in der α1-Helix, als auch am Ende dieser Helix es zu einem markanten Energieabfall gekommen ist. Die BRCT-Domänen aus TopBP1 0/2-M und TopBP1 7/8-N, welche sich ebenfalls in diesem C2 Cluster befinden, weisen hingegen eine vollkommen aktive Bindungstasche auf, da deren Energien sich in diesem Bereich auf einem hohen Niveau befinden. Anhand dieser energetischen Analysen ist es nun möglich, dass konkrete Aussagen über den evolutionären Verlauf bzw. über die Aufteilung der Taxa getroffen werden können.

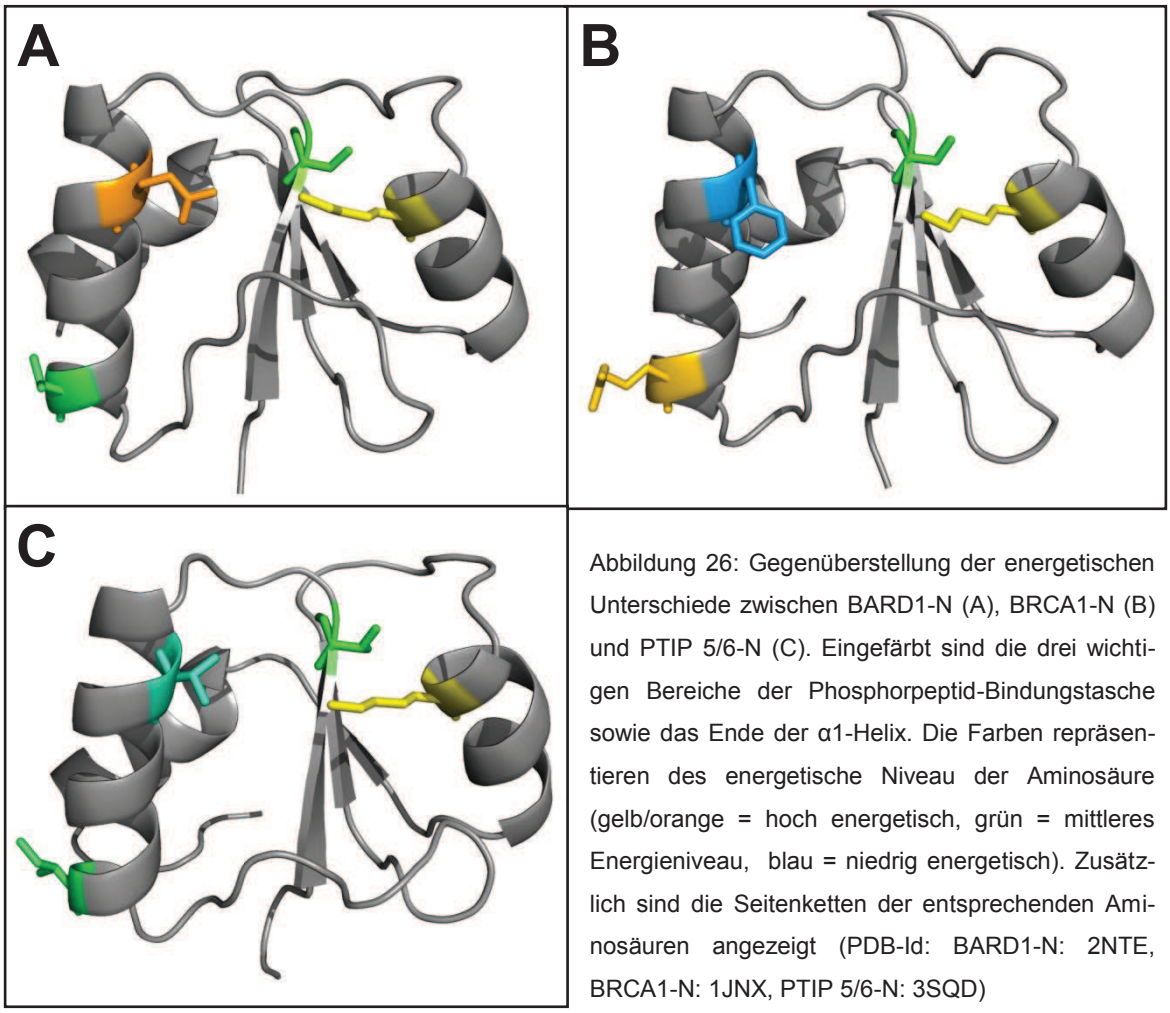


Abbildung 26: Gegenüberstellung der energetischen Unterschiede zwischen BARD1-N (A), BRCA1-N (B) und PTIP 5/6-N (C). Eingefärbt sind die drei wichtigen Bereiche der Phosphorpeptid-Bindungstasche sowie das Ende der α1-Helix. Die Farben repräsentieren des energetische Niveau der Aminosäure (gelb/orange = hoch energetisch, grün = mittleres Energieniveau, blau = niedrig energetisch). Zusätzlich sind die Seitenketten der entsprechenden Aminosäuren angezeigt (PDB-Id: BARD1-N: 2NTE, BRCA1-N: 1JNX, PTIP 5/6-N: 3SQD)

Jedoch ist es wichtig zu erwähnen, dass der Hauptgrund für die Aufteilung der beiden Cluster auf die sehr hohe strukturelle Ähnlichkeit der Domänen zurück zuführen ist. Somit bilden die Domänen aus C1 einen Cluster da deren Strukturen beinahe identisch erscheinen. Jedoch gibt es zwei energetische Merkmale, welche für die genauere Aufteilung verantwortlich sein könnten. Dies wären zum einen der Energieabfall am Ende der α1-Helix bei BARD1-N und die starke energetische Veränderung der Bindungstasche bei BRCA1-N. In dem C2 Cluster spalten sich TopBP1 0/2-M und TopBP1 7/8-N von den restlichen Taxa ab, da sie sowohl auf struktureller als auch energetischer Ebene eine sehr hohe Ähnlichkeit zueinander aufweisen. PTIP 5/6-N spaltet sich vermutlich aufgrund der zwei energetischen Veränderungen innerhalb der α1-Helix ab. Die Einreihung der Dlig IV-C Domäne ist jedoch strukturell zu erklären. Wie festzustellen ist, befinden sich mit Ausnahme der Dlig IV-C Domäne in diesen beiden Clustern nur N-terminale Domänen, welche eine aktive Phosphorbindetasche aufweisen. Bei den meisten dazugehörigen C-terminalen Domänen ist diese Bindungstasche nicht vorhanden und sie weisen, im Vergleich zu den N-terminalen Domänen eine stark veränderte Struktur auf. Die Dlig IV-C

Domäne verfügt ebenfalls nicht über diese Bindungstasche, jedoch weist sie strukturell gesehen eindeutig eine höhere Ähnlichkeit zu den N-terminalen d1 Domänen auf als zu den C-terminalen. Dies kann evtl. daran liegen, dass die Domänen der d2 Gruppe eine andere proteinbindende Funktion aufweisen als die Domänen der d1 Gruppe, jedoch kann dies, wie öfters erwähnt, aufgrund von Datenmangel der d2 Domänen nicht näher untersucht werden. Bei der energetischen Analyse dieser beiden Cluster zeigte sich jedoch nun, dass anhand dieser eine plausible Erklärung für die Aufteilung der Domänen in diesem Cluster gefunden werden konnte. Versucht man vergleichsweise die Aufspaltungskriterien dieser Cluster in dem NJ-Struktur Baum zu identifizieren so zeigt sich, dass sich auf struktureller Ebene nur ungenaue und weniger plausible Aussagen über die Aufteilung der Taxa getroffen werden können. So wäre z.B. nur schwer die Aufteilung zwischen BRCA1-N, BARD1-N und MCPH1 2/3-N nachvollziehbar. Strukturell ist nämlich zu erkennen, dass der markanteste Unterschied zwischen diesen drei Domänen jener ist, dass MCPH1 2/3-N, im Gegensatz zu BARD1-N und BRCA1-N ein β2-Sheet besitzt. Logischerweise würde man anhand dieser Tatsache davon ausgehen, dass BARD1-N und BRCA1-N eher miteinander verwandt sind, da deren Struktur optisch etwas ähnlicher zueinander sind als zu MCPH1 2/3-N. Somit ist in dem NJ-Struktur Baum nicht ersichtlich, welche strukturellen Veränderungen zu der dargestellten Aufteilung geführt haben könnten.

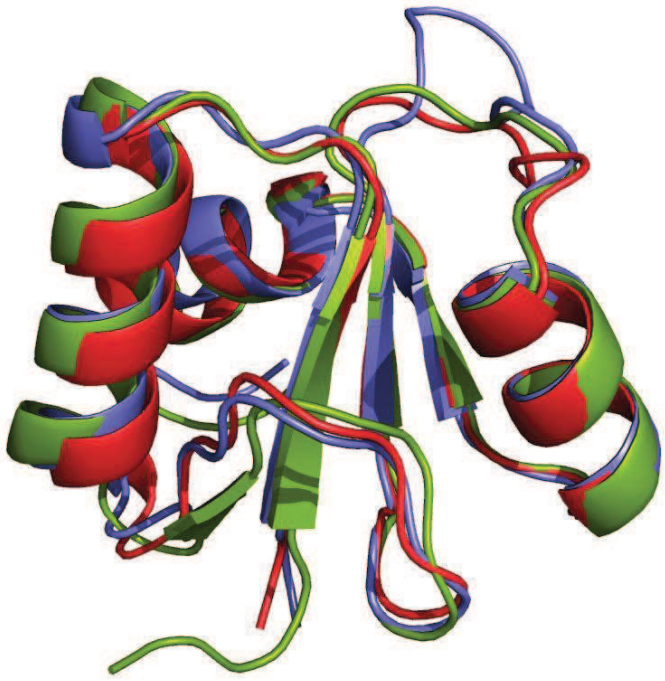


Abbildung 27: Strukturalignment von BARD1-N (rot), BRCA1-N (blau) und MCPH1 2/3-N (grün). (PDB-Id: BARD1-N: 2NTE, BRCA1-N: 1JNX, MCPH1 2/3-N: 3T1N)

Des Weiteren sind nur schwer die Auswirkungen auf die funktionellen Eigenschaften nachvollziehbar. So kann nicht ausgesagt werden, in wie fern z.B. das fehlende β2-Sheet in BARD1-N und BRCA1-N Auswirkungen auf die Funktion der Domäne hat. Diese Nachteile der rein strukturellen Betrachtung sind jedoch auf energetischer Ebene nicht vorhanden. So zeigt sich, dass das Fehlen des β2-Sheet keinerlei Auswirkungen hat, da in allen drei Domänen an dieser Stelle ein gleiches energetisches Niveau vorhanden ist. Somit ist das Vorhandensein eines β-Sheet an dieser Stelle nicht wichtig, sonder es ist wichtig, dass ein bestimmtes energetisches Verhalten vorhanden ist. Des Weiteren wäre auf struktureller Ebene nicht ersichtlich, dass in BRCA1-N eine Veränderung der Bindungstasche erfolgte. Zwar ist es möglich, mit Hilfe des PyMol Programmes sich die entsprechenden Stellen anzuschauen und sich auch die Seitenketten anzeigen zu lassen, welche für das Ausbilden der Bindungen wichtig sind, jedoch kann nicht ausgesagt werden, in wie weit nun diese wirklich eine Bindung zulassen. Es ist bekannt, dass besonders oft Lysin, Arginin oder Glutamin in diesen Bereich anzutreffen sind, da sie das Ausbilden einer Bindung gut ermöglichen (siehe Punkt 4), jedoch ist es bei dem Vorhandensein einer anderen Aminosäure schwierig auszusagen ob diese nun eine positive oder negative Auswirkung hat. Auf energetischer Ebene ist jedoch gut und schnell zu erkennen, dass das Auftreten negativer Energien an einer dieser Stellen sich negativ auf die Bindungsaktivität auswirkt.

Die Analyse dieser beiden Cluster zeigt nun, dass bevorzugt bei den N-terminalen Domänen der BRCT-Domäne eine aktive Phosphorbindungstasche auftritt. Des Weiteren zeigte die Analyse, dass trotz hoher struktureller Ähnlichkeit es klare energetische Differenzen gibt, welche zu einer Aufteilung der Domänen führte. Dennoch ist die strukturelle Ähnlichkeit der gröbere Faktor, anhand dessen die Aufteilung in die Cluster erfolgte. In wie weit nun die strukturelle Ähnlichkeit diese Aufteilung beeinflusst, war bei der Analyse der nächsten beiden Cluster ersichtlicher.

8.2 Analyse der C-terminalen Double BRCT-Domänen

Die nächsten Cluster, welche einer genaueren Analyse unterzogen wurden, waren die C3 und C4 Cluster. Wie auf den ersten Blick festzustellen ist, befinden sich in diesen hauptsächlich die C-terminalen Domänen der soeben untersuchten N-terminalen Domänen. Als erstes erfolgte nun die Betrachtung der Domänen des C3 Clusters durch das entsprechende MEPAL. Wie sich jedoch zeigte, waren aus diesem und dem MEPAL des C4 Clusters keine genauen bzw. markanten Veränderungen ersichtlich. Bei der Betrachtung

durch das Strukturalignment zeigte sich auch der Grund dafür. Im Gegensatz zu den N-terminalen Domänen herrscht bei den C-terminalen eine weitaus größere strukturellere Vielfalt. Besonders die Regionen zwischen der α1-Helix und dem β3-Sheet, sowie zwischen dem β3-Sheet der α3-Helix sind von deutlichen strukturellen Veränderungen durchzogen. So zeigt sich, dass in der Region vor dem β3-Sheet alle Domänen des C3 Cluster eine zusätzliche α-Helix besitzen, welche jedoch unterschiedlich stark ausgeprägt ist. Des Weiteren ist strukturell zu erkennen, dass je nach Domäne unterschiedlich viele β-Sheet auftreten. So verfügen die 53bp1-C und TopBP1 0/2-N Domänen über ein markantes und großes β2-Sheet, jedoch ist kein β4-Sheet vorhanden. In BARD1-C jedoch ist dies umgekehrt, da dort kein β2-Sheet aber ein deutliches β4-Sheet vorhanden ist. Die PTIP 5/6-C Domäne wiederum weist weder ein β2-Sheet noch ein β4-Sheet auf.Betrachtet man nun diese Bereiche des β2- und β4-Sheets auf energetischer Ebene so zeigt sich, dass bei allen vier Domänen an diesen Stellen ein sehr ähnliches energetisches Niveau vorhanden ist. Dieses Niveau ist zudem mit jenen aus den N-terminalen Domänen der C1 und C2 Cluster vergleichbar. Somit zeigt sich auch in diesem Cluster, dass das Vorhandensein eines β2-Sheets nicht zwingend notwendig ist, sondern lediglich ein entsprechendes energetisches Niveau benötigt wird.

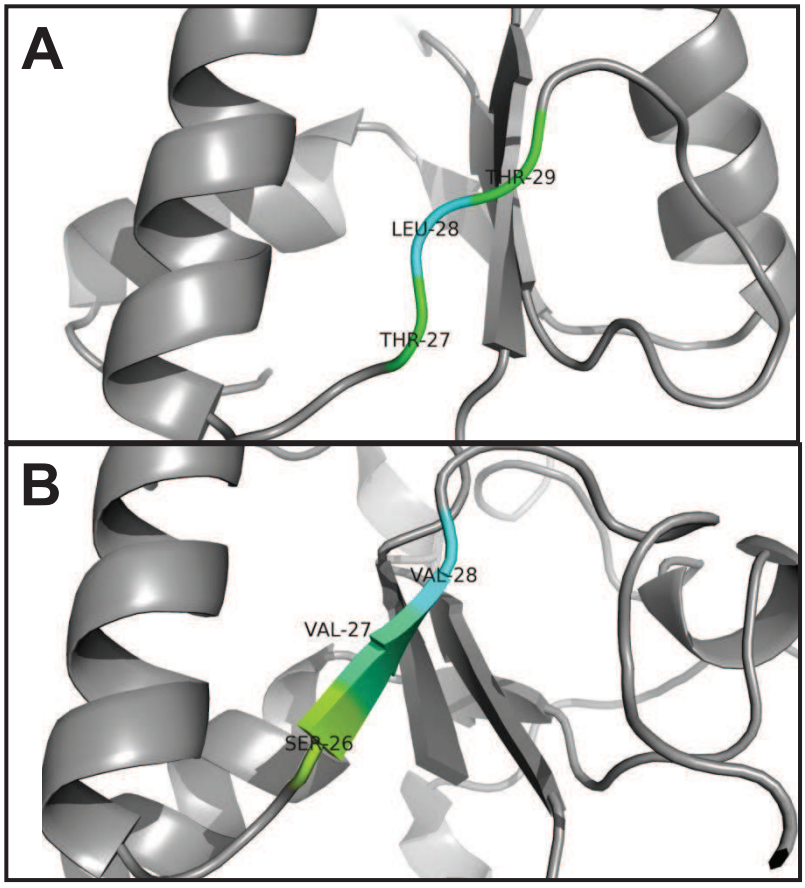


Abbildung 28: Vergleich der Energien an der Position des β2-Sheets in BRCA1-N (A) und BRCA1-C (B). An den energetisch eingefärbten Positionen sind die entsprechenden Aminosäuren angegeben. (PDB-Id: 2NTE)

Diese Tatsache lässt sich auch auf das β4-Sheet erweitern. Während dieses bei den N-terminalen Domänen noch zum Großteil vorhanden ist, scheint es bei den C-terminalen Domänen keine wichtige Rolle zu spielen. Vergleicht man auch an dieser Stelle die Energien aller vier bisher untersuchten Cluster miteinander, so zeigt sich, dass auch an dieser Stelle immer ein gleichbleibendes energetisches Niveau zu erkennen ist und somit das Ausbilden eines β4-Sheet nichts zwangsläufig notwendig ist.

Durch die große strukturelle Vielfalt dieser Domänen kommt es ebenfalls zu einer Veränderung der Phosphorpeptid-Bindungstasche. Die auffälligsten Veränderungen sind dabei zum einen der Wegfall der Seitenketten aus der α2-Helix, da diese starken strukturellen und räumlichen Veränderungen unterliegt. So ist diese beispielsweise in BARD1-C komplett verschwunden und wurde durch einen langen Coilbereich ersetzt. In TopBP1 0/2-N fand stattdessen eine räumliche Verschiebung dieser statt, so dass deren Seitenkette keinen Einfluss mehr auf die Bindungstasche nehmen können. Die zweite auffälligste Veränderung ist die Blockade der Bindungstasche, so wie es beispielsweise in 53bp1-C durch das β2-Sheet geschieht. Jedoch sind diese strukturellen Veränderungen nicht die einzigen, welche zu einem Verlust der Aktivität der Bindungstasche geführt haben. So zeigt sich auf energetischer Ebene, dass bei den Bereichen der α1-Helix und dem β1-Sheet der Bindungstasche, deutliche energetische Veränderungen stattfanden. Strukturell sind bei einigen Domänen, wie z.B. BARD1-C und PTIP 2/3-C diese beiden Stellen unverändert, was darauf hindeuten könnte, dass noch ein Teil der Bindungsaktivität vorhanden ist. Betrachtet man diese beiden Domänen jedoch energetisch, so zeigt sich, dass in beiden Domänen die Bindungsaktivität des β1-Sheet nicht mehr vorhanden ist, da an dieser Stelle starke negative Energien auftreten. Ähnlich verhält es sich an der Bindungsstelle in der α1-Helix. Während bei BARD1-C noch eine hohe Energie zu erkennen ist, weist PTIP 2/3-C an dieser Stelle bereits eine stark niedrige Energie auf. Betrachtet man die beiden Aminosäuren an dieser Stelle genauer (Lysin in BARD1-C und Leucin in PTIP 2/3-C), so zeigt sich, dass beide eine recht ähnliche Seitenkette besitzen, welche räumlich gleich ausgerichtet ist. Somit könnte man auch bei diesen beiden Domänen anhand struktureller Merkmale keine Aussage treffen, in wie weit eine Aktivität der Bindungstasche vorhanden ist. Energetisch ist jedoch erkennbar, dass BARD1-C noch über Reste dieser verfügt. Ein ähnliches energetisches Verhalten an dieser Stelle findet sich zudem auch bei den C-terminalen Domänen des C4 Clusters. So weisen TopBP1 7/8-C und BRCA1-C bei dieser Stelle in der α1-Helix ebenfalls noch erhöhte Energien auf. Eine anfängliche Annahme war es nun, dass diese erhöhten Energien eine Bedeutung für die proteinstabilisierende Funktion besitzt. Da jedoch die TopBP1 7/8-C Domäne bereits im Vorfeld einer

genauen Untersuchung unterzogen wurde, ist deren Funktion und die dafür beteiligten Aminosäuren weitestgehend klar (siehe Punkt 4.6). So stellt sich heraus, dass dieser Bereich der α1-Helix keinerlei Einfluss auf diese Bindungseigenschaft nimmt. Da jedoch das Auftreten dieser hohen Energien an dieser wichtigen Position nicht als Zufall betrachtet werden soll, lässt sich nun eine neue Annahme bzw. Hypothese aufstellen. Diese besagt, dass es sich bei den energetischen Bereichen um evolutionäre Überreste handelt. Es wird somit angenommen, dass sich die C-terminalen Domänen durch Duplikation der N-terminalen entwickelt haben und sich entsprechend die Bindungsaktivität der Bindungstasche veränderte. Dies führte unter anderem zu einem starken Energieabfall am Ende des β1-Sheets, da dieses dadurch an Stabilität gewann. Zum anderen mussten die Bereiche zwischen der α1-Helix und dem β3-Sheet, sowie zwischen dem β3-Sheet der α3-Helix, inklusive der α2-Helix keine tragende Rolle mehr für das aufbringen einer Proteinbindung spielen, was zu größeren strukturellen Veränderungen führte. Ebenfalls musste der Bereich innerhalb der α1-Helix keine bindende Funktion mehr erfühlen und passte sich energetisch den Veränderungen der Struktur an. Bei Strukturen wo jedoch in diesem Bereich nur wenige Veränderungen stattgefunden haben, erfolgten ebenfalls keine großen Veränderungen des energetischen Verhaltens der Aminosäuren. Somit kann logisch nachvollzogen werden, wieso sich bei einigen C-terminalen Domänen noch immer hohe Energien in diesem Bereich der Bindungstasche finden lassen. Des Weiteren können diese als Beweis für die soeben aufgestellte Annahme dienen, welche wiederum einen wichtigen evolutionären Schritt der BRCT-Domäne erklären würde. Versucht man nun eine plausible Begründung für die Beschaffenheit und Einteilung dieser zwei Cluster zu finden, so ist diese gleichermaßen auf struktureller und energetischer Ebene zu finden. Die Domänen des C3 Clusters scheinen sich eindeutig aufgrund ihrer strukturellen Eigenschaften hin zusammengefunden zu haben. Des Weiteren spielt dabei der Verlust der Phosphorbindetasche auf energetischer Ebene eine wichtige Rolle. Bei den Domänen des C4 Clusters sind diese beiden Punkte ebenfalls der Grund für die Darstellung der Verwandtschaft zueinander. Da 53bp1-N über eine aktive Bindungstasche verfügt, scheint es, dass die Einreihung von TopBP1 7/3-C und BRCA1-C aufgrund ihrer teilweise noch vorhandenen energetischen Reste der Bindungstasche und ihrer strukturellen Ähnlichkeit wegen erfolgte. BARD1-C verfügt zwar ebenfalls noch über diesen energetischen Rest, jedoch scheint die strukturelle Vielfalt zu groß, weswegen eine Einreihung in den C3 Cluster erfolgte. Des Weiteren scheinen ebenfalls die großen strukturellen Unterschiede von 53bp1-N zu den anderen N-terminalen Domänen der Grund zu sein, weswegen diese keine Einordnung in den C1 oder C2 Cluster findet. Eine genauere Erklärung für die Beschaffenheit und Aufteilung der Cluster kann jedoch nicht gegeben werden, da im Gegensatz zu den Domänen

des C1 und C2 Cluster bei diesen eine zu hohe Varianz auf struktureller Ebene und daraus folgend auch auf energetischer Ebene vorhanden ist. Dennoch lässt sich plausibel die Aufteilung der C-terminalen von den N-terminalen Domänen erklären.

8.3 Analyse der Single BRCT-Domänen

Nachdem nun die Domänen der d1-Gruppen ausführlich auf ihre evolutionären Verhältnisse und Eigenschaften hin untersucht wurden, sollen nun die Domänen der s2-Gruppe einer genaueren Analyse unterzogen werden. Dabei soll mit den Domänen des C5 Cluster begonnen werden. Wie auch bei den Domänen des C1 Clusters sind in diesem C5 Cluster die TDT und Dpoly Mu Domänen in allen Bäumen der drei unterschiedlichen Grundlagen anzutreffen. Dies lässt ebenfalls darauf schließen, dass diese eine besonders markante Eigenschaft aufweisen, ähnlich wie es bei den Domänen des C1 Cluster durch die sehr hohe strukturelle Ähnlichkeit gegeben ist. Betrachtet man doch zunächst das MEPAL dieses Clusters, so zeigt sich, dass TDT und Dpoly Mu zwei auffällige energetische Veränderungen, im Vergleich zu den anderen drei Domänen dieses Clusters aufweisen. Eine dieser Stellen findet sich in dem kurzen Coil Bereich, welcher sich zwischen der α1-Helix und dem β2-Sheet befindet (auch als GG-Loop bezeichnet, siehe Punkt 4.2). Die zweite Stelle befindet sich kurz davor am Ende der α1-Helix. Dieser Bereich am Ende der α1-Helix weist auch bei den d1 Domänen des C1 und C2 Clusters teilweise unterschiedliche Energiewerte auf (entweder deutlich hohe oder niedrige Werte). Bei TDT und Dpoly Mu finden sich an dieser Stelle deutlich hohe Energien, während bei den anderen Domänen dieses C5 Clusters (REV1, PES1, Dpoly lamda) deutlich niedrige Energien auftreten. Welche Bedeutung diese energetischen Differenzen an dieser Stelle aufweisen und welcher Bezug zu den jeweiligen Domänen der unterschiedlichen evolutionären Gruppen dadurch vorhanden ist, kann jedoch momentan noch nicht genau nachvollzogen werden. Jedoch scheint eine gewisse Systematik hinter diesem energetischen Verhalten zu stecken. Über die energetischen Veränderungen des GG-Loops lässt sich hingegen eine etwas genauere Aussage treffen. Wie bereits im Punkt 4.2 dargelegt, ist dieser kurze Loop Bereich der wohl wichtigste Bereich in allen BRCT Domänen. Bisherige sequenzielle Analysen zeigten, dass in diesem Bereich besonders häufig Glycin und Alanin anzutreffen sind. Diese Tatsache konnte bisher untermauert werden, da in fast allen bisher untersuchten BRCT-Domänen diese beiden Aminosäuren an besagter Stelle ebenfalls anzutreffen sind. Die einzige Ausnahme bilden dabei die drei Domänen des C1 Clusters sowie TDT

und Dpoly Mu, in welchen zum Teil auch andere Aminosäuren entsprechend vorhanden sind.

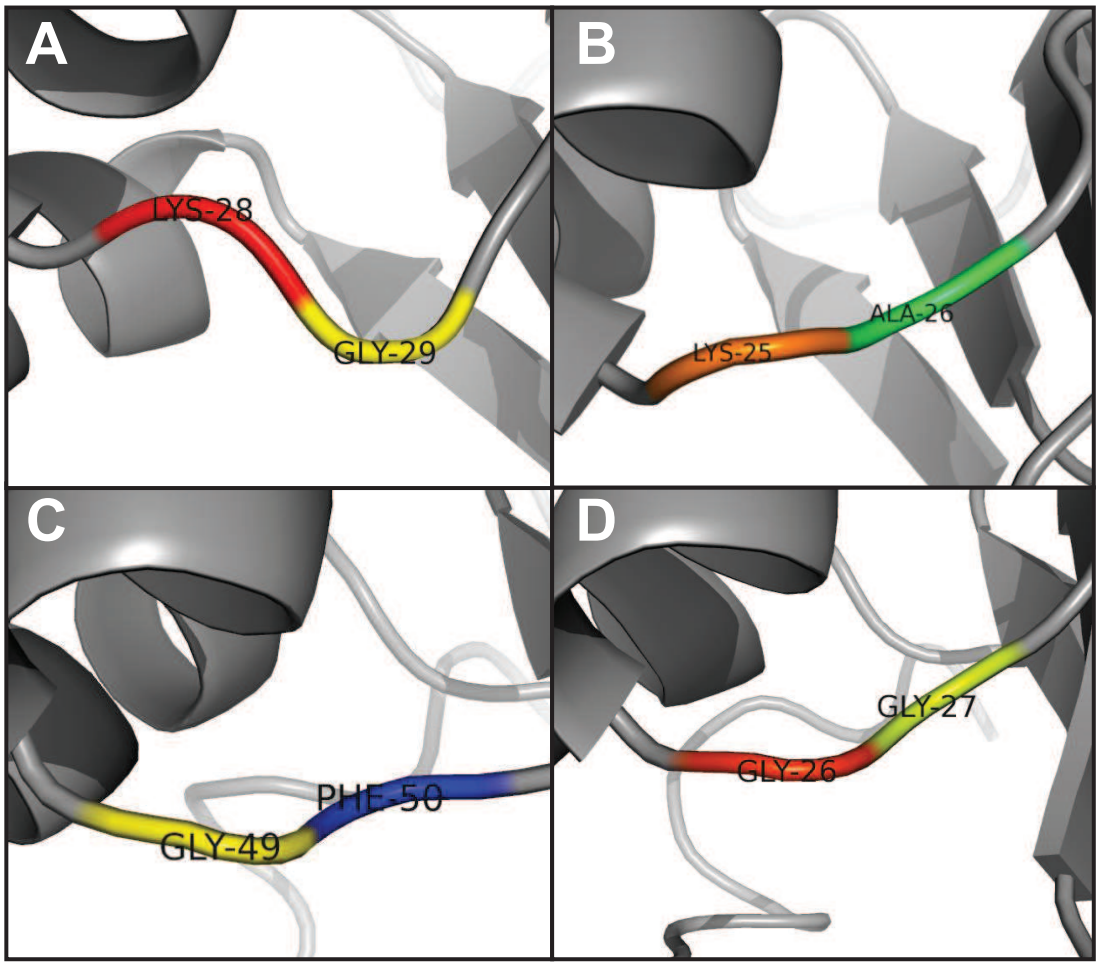


Abbildung 29: Gegenüberstellung der Energien des GG-Loops aus vier verschiedenen BRCT-Domänen. Wie sich zeigt besitzt die Erste Aminosäure des Loops immer eine höhere Energie als die Zweite. Die Anwesenheit von Glycin (Gly) oder Alanin (Ala) ist nicht zwingen notwendig! A: MCPH1 2/3-N (PDB-Id: 3T1N), B: BARD1-N (PDB-Id: 2NTE), C: TDT (PDB-Id: 2COE), D: PTIP 5/6-C (PDB-Id: 3SQD)

Laut bisherigen Untersuchungen auf sequenzieller und struktureller Ebene würden theoretisch diese sequenzielle Veränderungen an dieser Stelle zu einer Destabilisierung der gesamten Struktur führen. Jedoch kann dies teilweise wiederlegt werden, indem auf die Energiewerte dieser zwei Aminosäuren an jener Stelle genauer eingegangen wird. So zeigt sich in allen Strukturen der BRCT-Domänen, dass an dieser Stelle, unabhängig von der vorkommenden Aminosäure, immer die erste Aminosäure eine deutlich höhere Energie aufweist als die zweite und anschließend ein gleichbleibendes energetisches Niveau gehalten wird an jener Stelle wo sich gegebenenfalls das β2-Sheet befindet. So haben sich Glycin und Arginin aufgrund ihrer Eigenschaften an dieser Stelle etabliert, jedoch ist

anscheinend eine Stabilität der Struktur auch noch dann gewährleistet, wenn an dieser Stelle ein gleichbleibendes energetisches Verhalten herrscht. So führen Mutationen in diesem Loop Bereich vermutlich nur dann zu negativen Veränderungen, wenn sich etwas an dieser fest vorgeschriebenen energetischen Abfolge ändert. Diese energetische Veränderung des Loop Bereiches in TDT und Dpoly Mu, welche jedoch keine negativen Einfluss auf die Stabilität der Struktur aufweist, sowie die energetische Veränderung am Ende der α1-Helix sind der Grund dafür, dass eine Abspaltung dieser beiden Taxa innerhalb des C5 Clusters erfolgte. Jedoch sind sie vermutlich nicht der maßgebliche Grund für die hohe Konservierung innerhalb der Bäume der anderen Grundlagen. Der Hauptgrund dafür liegt mit hoher Wahrscheinlichkeit ebenfalls strukturell begründet. So zeigt sich im Strukturalignment, dass die α2-Helix in TDT und Dpoly Mu, sowie auch Dpoly lamda eine besondere räumliche Ausrichtung aufweist. Entgegen der α1- und α3-Helix, so wie alle anderen vorkommenden α2-Helix in den anderen Domänen weist diese in diesen drei Domänen eine um ca. 90° in der Vertikalen gekippte Ausrichtung auf. Der Grund für diese deutlich veränderte Ausrichtung ist jedoch unklar. Da alle Domänen der s2 Gruppen keine aktive Phosphorbindetasche aufweisen, kann man diese besondere Ausrichtung als evolutionären Versuch deuten, die Bindeaktivität der Tasche negativ zu beeinflussen. Jedoch zeigte sich, dass durch diese räumliche Veränderung keine gravierenden Veränderungen der Bindungstasche erfolgten und somit theoretisch immer noch die Bindung eines phosphoreszierten Peptides möglich wäre.

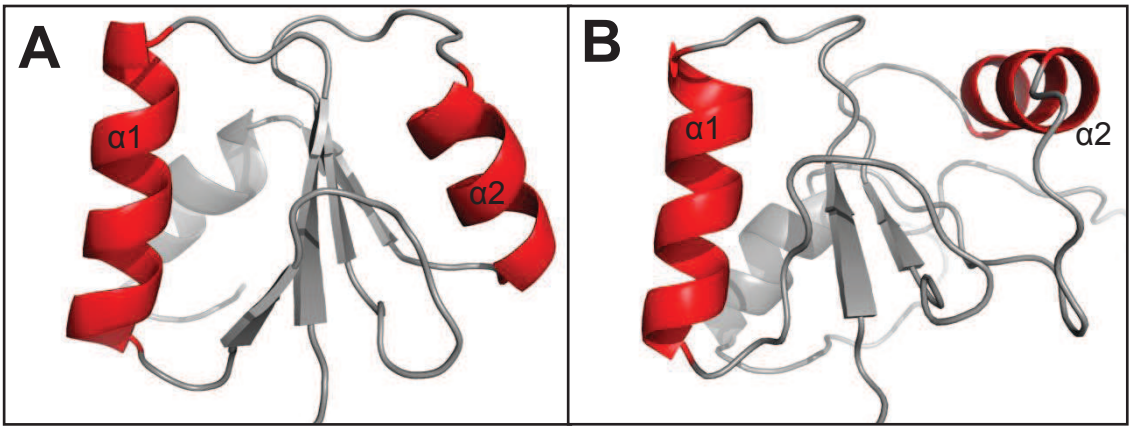


Abbildung 30: Vergleich der räumlichen Ausrichtung der α2-Helix. A: Typische Ausrichtung der α2-Helix wie sie in den meisten BRCT-Domänen anzutreffen ist, anhand des Beispiels von MCPH1 2/3-N (PDB-Id: 3T1N). Dabei sind die α1- und α2-Helix rot eingefärbt. B: Abbildung der um etwa 90° gekippte α2-Helix in TDT (PDB-Id: 2COE).

Auch die genauere Betrachtung der entsprechenden Seitenketten in dieser Bindungstasche lässt diesen Schluss zu, welche noch dahingehend gestärkt wird, dass in einigen der Domänen sich noch typische Peptidbinder in der α1-Helix finden lassen. Jedoch zeigt sich abermals auf energetischer Ebene, dass sowohl die Seitenketten der α2-Helix als auch jene des β1-Sheetes aufgrund ihrer starken negativen Energien keine Bindung eines Peptides zulassen. Auch die typischen Peptidbinder der α1-Helix, wie z.B. Arginin in TDT und PES1 weisen im Vergleich zu jenen Aminosäuren einer aktiven Bindungstasche deutlich niedrigere Energien auf. Dieses energetische Verhalten an dieser Stelle der Bindungstasche erinnert jedoch stark an jenes, welches bei einigen der C-terminalen Domänen der d1 Gruppe in den C3 und C4 Clusters beobachtet werden konnte. Generell bestehet zwischen den C-terminalen Domänen aus C3 und C4 und den Domänen aus C5 eine signifikante strukturelle Ähnlichkeit, welche auf eine gemeinsame evolutionäre Entwicklung hindeutet. Somit könnte die Hypothese über den evolutionären Verlauf, welche vorhin aufgestellt wurde, dahingehen erweitert werden, dass eine Entwicklung der Domänen der s2 Gruppe parallel zu den C-terminalen Domänen der d1 erfolgte. Etwas erweiter würde dies bedeuten, dass sich zuerst die N-terminalen Domänen der d1 Gruppe aus den Domänen der s1 Gruppe entwickelt haben und sich später die C-terminalen durch Duplikation dieser entwickelten, während einige N-terminale weiterhin einzeln auftretend blieben, sich jedoch in diesen die gleichen funktionellen Veränderungen wie bei den C-termianlen Domänen vollzogen haben. Vor allem die deutlichen Ähnlichkeiten zwischen den C-terminalen Domänen und jenen der s2 Gruppe in C5 sorgen in den Sequenz- und Strukturbäumen dafür, dass in diesen eine Vermischung der Taxa erfolgt. Da eine derartig intensive Vermischung, trotz struktureller Ähnlichkeiten und energetischen Ähnlichkeiten innerhalb der Phosphorbindungstasche, in den EP Bäumen nicht erfolgt, ist anzunehmen, dass zwischen den Domänen dieser Gruppen eine eindeutige energetische Differenz vorhanden sein könnte, welche für die Aufspaltung der evolutionären Gruppen sorgt. Die Suche nach diesem eindeutigen energetischen Unterschied erwies sich jedoch als schwierig. Als Anhaltspunkt für die Suche nach dieser Differenz dienten die funktionellen Eigenschaften der Domänen. So ist es die Aufgabe der s2 Domänen eine Bindung bevorzugt zu einer anderen BRCT-Domäne über die jeweilige α1-Helix und teilweise auch α3-Helix auszubilden. Diese Funktion legte die Annahme nahe, dass eine gewisse energetische Konservierung in diesen Bereichen zu erkennen ist. Jedoch zeigt sich, dass sowohl die C-terminalen als auch die N-terminalen Domänen der d1 und d2 Gruppe in diesen beiden Helices eine der s2 Gruppen entsprechende Energie aufweisen. Dies ist jedoch auch nicht weiter verwunderlich. Zum einen ist es laut der Theorie des hydrophoben Kollapses logisch, dass in einem Protein die Aminosäuren, welche zu dem Lösungsmittel zeigen, hohe energetische

Werte aufweisen (siehe Punkt 5). Zudem lassen sich bei den C-terminalen Domänen diese Energien dadurch erklären, da diese mit ihrer jeweiligen N-terminalen Domäne in dem Protein ein Dimer bildet, weswegen es an der Trennungsstelle durch den Linker dennoch zu einer Vielzahl an Wechselwirkungen zwischen der α1- und α3-Helix der C-terminalen und der α2-Helix der N-terminalen Domäne kommt. Von daher lassen sich auf energetischer Ebene nur sehr schwer evolutionäre Rückschlüsse durch diese Bereiche ziehen. Deshalb lässt an dieser Stelle nicht genau sagen, welcher Faktor zu einer deutlichen Trennung zwischen den Domänen der s2 Gruppe und der d1 Gruppe geführt haben könnte, da durch die hohen strukturellen Veränderungen entsprechend auch größere energetische Unterschiede vorliegen. Auch die Analyse des nächsten Cluster konnte keinen Aufschluss darüber geben, welcher Faktor für diese Trennung verantwortlich erscheint.

Der vorletzte Cluster, welcher genauer untersucht wurde, ist der an die C1/C2 Cluster anbindente C6 Cluster. Entgegen der s2 Domänen des C5 Clusters, scheinen die Domänen dieses Clusters eher zu den N-terminalen Domänen der d1 Gruppe Ähnlichkeiten aufzuweisen. Dies wird schon dadurch bewiesen, dass sich die TopBP1 0/2-C Domäne in diesen Cluster eingeordnet hat. Entgegen aller anderen C-terminalen Domänen besitzt TopBP1 0/2-C eine aktive Phosphorpeptid Bindungstasche, so wie es bereits schon in früheren Studien festgestellt wurde (siehe Punkt 4.6). Von daher besitzt diese eine sehr hohe energetische Ähnlichkeit zu der TopBP1 0/2-M Domäne, welche ebenfalls über diese Bindungstasche verfügt und ist dementsprechend auch in der Nähe der C1 und C2 Cluster angesiedelt. Die Tatsache, dass sich diese nun zwischen einigen Domänen der s2 Gruppe angelagert hat impliziert die Vermutung, dass diese s2 Domänen energetisch gesehen noch über markante Reste der Bindungstasche verfügen müssen. Betrachtet man das energetisch eingefärbte Strukturalignment so zeigt sich deutlich, dass sich diese Vermutung als wahr erweist. Denn mit Ausnahme von MCPH1-s verfügen die anderen drei s2 Domänen dieses Clusters über sehr deutliche hochenergetische Reste innerhalb der Bindungstasche. Betrachtet man diese Reste genauer, so zeigt sich, dass lediglich in dem β1-Sheet ein teilweise ersichtlicher Rückgang der Bindungsaktivität zu erkennen ist. In der α1- und α2-Helix sind jedoch noch die Bindungsbereiche sowohl energetisch als auch sequenziell in einer hohen Konservierung vorhanden, so wie man sie auch in den Domänen der s1 Gruppe und den meisten N-terminalen Domänen der d1 Gruppe findet. Über den Grund dieser hohen Konservierung kann jedoch nur spekuliert werden. Eine mögliche Theorie, welche als Erklärung für dieses Verhalten dienen könnte, wäre, dass sich diese s2 Domänen erst vor einem relativ kurzen Zeitraum aus den s1 (oder auch den N-terminalen d1) Domänen entwickelt haben, da sich bisher lediglich an einer energeti-

schen Stelle eine markante evolutionäre Veränderung vollzogen hat, während die restlichen energetischen und strukturellen Merkmale noch denen der N-terminalen d1 Domänen sehr ähnlich sind. Daraus kann angenommen werden, dass sich diese Domänen am Anfang eines Veränderungszyklus befinden, bei welchem sich die Bindungstasche zurück entwickelt. Folglich befinden sich die s2 Domänen des C5 Clusters in einem fortgeschrittenem Stadium dieses Veränderungszykluses, da zwar noch energetische Merkmal vorhanden sind, sich jedoch schon deutliche strukturelle Veränderungen vollzogen haben. Die MCPH1-s Domäne scheint somit theoretisch und evolutionär gesehen zwischen diesen beiden Clustern zu stehen, da deren Bindungstasche deutlichen energetischen Veränderungen unterlag, aber strukturell gesehen noch signifikante Ähnlichkeiten zu den C1, C2 und C6 Domänen besitzt.

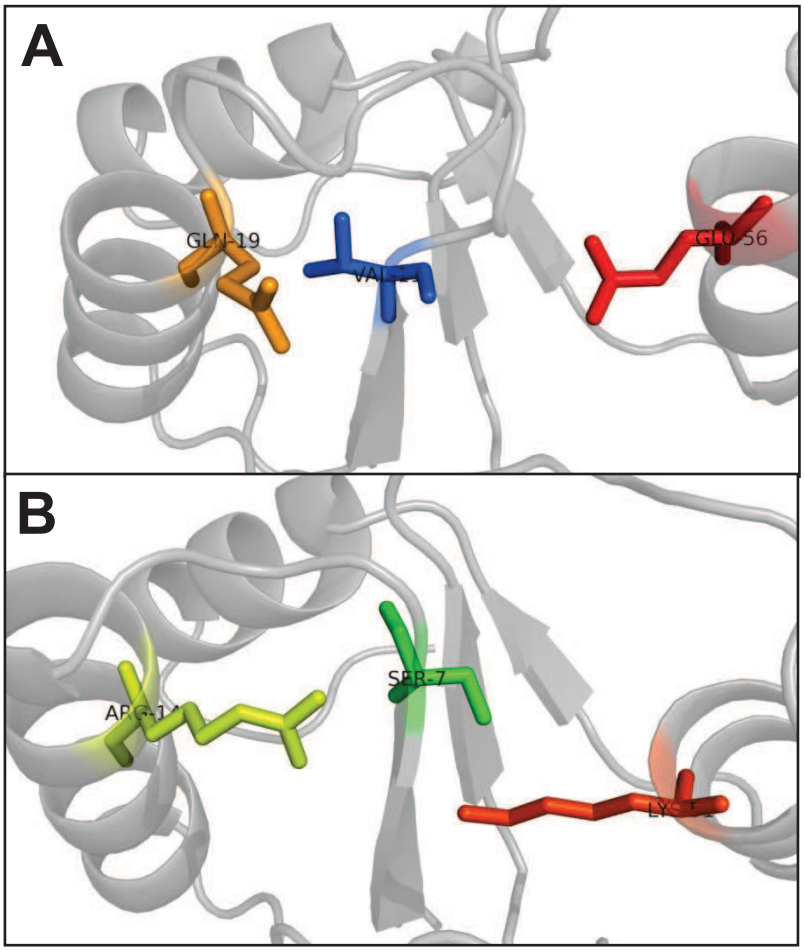


Abbildung 31: Vergleich der Energien der Bindungstasche einer N-terminalen d1 Domäne mit einer s2 Domäne aus dem C6 Cluster. A: Energetisch eingefärbte Aminosäuren der Bindungstasche aus TopBP1 6 (PDB-Id: 3PD7), welche in die s2 Gruppe eingeordnet ist und sich im C6 Cluster befindet. B: Energien der Bindungstasche von TopBP1 7/8-N (PDB-Id: 3AL2)aus dem C2 Cluster.

Eine andere Theorie, welche man jedoch ebenfalls in Betracht ziehen könnte wäre, dass die Einordnung dieser Domänen des C6 Clusters in die evolutionäre s2 Gruppe falsch ist. Es sind zwar energetische Veränderungen in dem β1-Sheet der Bindungstasche zu erkennen, jedoch sind diese teilweise nicht so deutlich, dass nun mit einer 100%igen Wahrscheinlichkeit gesagt werden kann, dass diese kein phosphoresziertes Peptid oder einen DNA Strang mehr binden. Um die wirkliche Bindungsaktivität feststellen zu können, müssten ausführliche Test und Simulationen durchgeführt werden, in denen das Binden eines entsprechenden Objekts an dieser Stelle simuliert wird. Erst danach kann genauer ausgesagt werden, in wie weit eine Bindungsaktivität vorhanden ist und ob diese Domänen in eine andere oder sogar neue evolutionäre Gruppe einzuordnen sind. Betrachtet man nun noch die restlichen bisher zur Analyse herangezogenen energetischen und strukturellen Merkmale, so zeigt sich zum einen, dass der GG-Loop keine besondere Veränderungen aufweist, sondern den weiter oben erklärten, für diesen Bereich typischen energetischen Verlauf aufweist, welcher in allen anderen Domänen bisher zu beobachten war. Auch die Bereiche des β2- und β4-Sheets weisen die bisher beobachteten und typischen energetischen Merkmale auf. Wie auch bei den anderen Clustern zu beobachten ist, verfügen auch einige Domänen in diesem Cluster nicht zwangsläufig über eines dieser beiden β-Sheets aber weisen dennoch ein entsprechendes energetisches Niveau auf.

Der nunmehr letzte Cluster, den es zu untersuchen galt, war der C7 Cluster, in dem Taxa aller vier evolutionären Gruppen zu finden sind. Da nun nach allen bisherigen Analysen der Beweis erbracht werden konnte, dass sich die Domänen anhand energetischer und struktureller Merkmale gut voneinander abtrennen und sich, je nach Stärke der Ausprägungen dieser Merkmale, in dem Stammbaum gut ersichtlich in Clustern anordnen, kann anhand dieses Wissen eine erste Annahme aufgestellt werden, welche die scheinbar willkürliche Anordnung aller vier evolutionären Gruppen in diesem Cluster näher erklärt. Diese Annahme ist die gleiche, wie sie ebenfalls bei den Domänen des C6 Clusters aufgestellt wurde, welche wiederrum besagt, dass es sich auch bei den Domänen dieses Clusters um evolutionäre Zwischenschritte handelt. Somit wären die Domänen der C1 bis C5 Cluster die bisherigen evolutionären „Endformen“ (die am besten an ihre Aufgaben angepassten Domänen) der verschiedenen evolutionären Gruppen während die Domänen des C6 und C7 Cluster sich noch in einem Veränderungszyklus befinden und teilweise noch nicht optimal an ihre jeweilige Aufgabe angepasst sind. Da sich die Domänen des C7 Clusters aber nun ebenfalls in einem deutlich separierten Cluster abgetrennt haben, sollte es theoretisch auch zwischen diesen ein vorherrschendes Merkmal geben, welches zu einer entsprechenden Einteilung geführt hat. Betrachtet man diese Domänen energetisch

im Strukturalignment, so zeigt sich, dass alle Domänen die unterschiedlichsten, bisher beobachteten Eigenschaften aufweisen. So verfügt MCPH1 2/3-C über keine aktive Bindungstasche, da sie sowohl die für C-terminalen typischen niedrigenergetischen Bereiche in dieser aufweist, als auch die ebenfalls typischen großen strukturellen Veränderungen in dem Bereich um die α2-Helix besitzt. Somit handelt es sich bei dieser um eine typische C-terminale Domäne, welche eher in den C4 oder C5 Cluster einzuordnen wäre. Bei FCP1, Dlig III und Dlig IV-C sind jedoch komplett andere energetische und strukturelle Verhältnisse zu erkennen. So weisen diese drei Domänen alle eine teilweise noch aktive Bindungstasche auf, so wie sie auch bei den s2 Domänen im C6 Cluster zu erkennen ist. Auch die strukturellen Merkmale dieser sind zu den Domänen des C6, C1 und C2 Cluster sehr ähnlich. Somit würden diese drei Domänen eher eine Einteilung in diese Cluster finden. Auch die Einordnung von RFC1 in diesen C7 Cluster ist sehr fraglich. Wie in den anderen Bäumen der drei verschiedenen Grundlagen zu erkennen ist, sind die verwandtschaftlichen Beziehungen der Domänen der s1 Gruppe nur sehr schwer nachzuvollziehen, da sie in allen Bäumen keine feste Zuordnung zu anderen Taxa finden. Der Grund dafür kann ebenfalls wie bei den Domänen der d2 Gruppe darin liegen, dass zu wenig Ausgangsdaten bzw. Strukturen von jenen vorhanden sind. Ein logisches Nachvollziehen, wieso RFC1 sich in diesem Cluster zu den anderen Domänen (besonders zu Dlig IV-C) eingeordnet hat, ist somit nicht möglich. Generell lässt sich anhand der bisher untersuchten Merkmale und dem Wissen, welches aus allen bisherigen Untersuchungen gezogen wurde, keine plausible Erklärung für Beschaffenheit des C7 Clusters finden. Die Ursache für die Existenz dieses Clusters könnte somit evtl. in der noch unausgereiften Anwendung von Energieprofilen auf phylogenetische Methoden bzw. deren Software liegen.

Am Ende sollen nun die Domänen der s1 Gruppe noch einer separaten Untersuchung unterzogen werden, da sich aus diesen bekanntermaßen die Domänen der s2, d1 und d2 Gruppe entwickelt haben sollen. Die Domänen der s1 Gruppe sind jene BRCT-Domänen, wie sie auch häufig in niederen Organismen, wie z.B. Bakterien anzutreffen sind. Von daher verkörpern diese die erste Funktion und Beschaffenheit, welche den BRCT-Domänen ursprünglich zugekommen war. Wie bereits bekannt ist, verfügen diese Domänen über eine ausgeprägte DNA bindende Funktion, welche durch Bindung des phosphoreszierten 5‘-ende des DNA Stranges erfolgt (siehe Punkt 4.7). Somit besitzen diese Domänen eine aktive Bindungstasche. Die Analyse dieser Bindungstasche mit Hilfe des energetisch gefärbten Strukturalignments zeigt, dass eine sehr starke Ausprägung hoher Energien in dieser zu finden ist. Diese sind teilweise sehr viel höher ausgeprägt als jene bei den N-terminalen Domänen des C1 und C2 Clusters. Zudem zeigt sich, dass in einem

zusätzlichen Bereich der α1-Helix ein stark hochenergetischer Bereich zu erkennen ist. Dieser Bereich befindet sich unmittelbar unterhalb des normalen Bindungsbereiches und liegt somit zentral in der Helix. Bei der Betrachtung der Seitenketten dieser Stelle zeigt sich, dass sie in etwa die gleiche Ausrichtung aufweisen wie jene am oberen Ende der Helix und somit Einfluss auf die Bindungstasche ausüben können. Dies legt die Annahme nahe, dass dieser zusätzliche hochenergetische Bereich für das Binden eines DNA Stranges eine wichtige Funktion hat. Betrachtet man diesen Bereich in den N-terminalen Domänen der d1 Gruppe, so zeigt sich, dass diese keine derartigen hohen Energien an jener Position aufweisen und somit davon aufgegangen werden kann, dass diese nicht für das Binden eines phosphoreszierten Peptides benötigt werden. Strukturell gesehen verfügt RFC1 und teilweise auch PARP1 über eine kleine zusätzliche Helix, welche oberhalb der Struktur angelagert ist. Diese dient zumindestens in RFC1 zur Stabilisierung des gebundenen DNA Stranges (siehe Punkt 4.3). Ansonsten besitzen sie, verglichen mit den N-terminalen Domänen, keine weiteren strukturellen oder energetischen Eigenheiten. Somit scheint es, dass sich die N-terminalen Domänen direkt und als erste aus den Domänen der s1 Gruppe entwickelt haben, da in diesen nur eine geringe Anpassung der energetischen Stellen der Bindungstasche erfolgte.

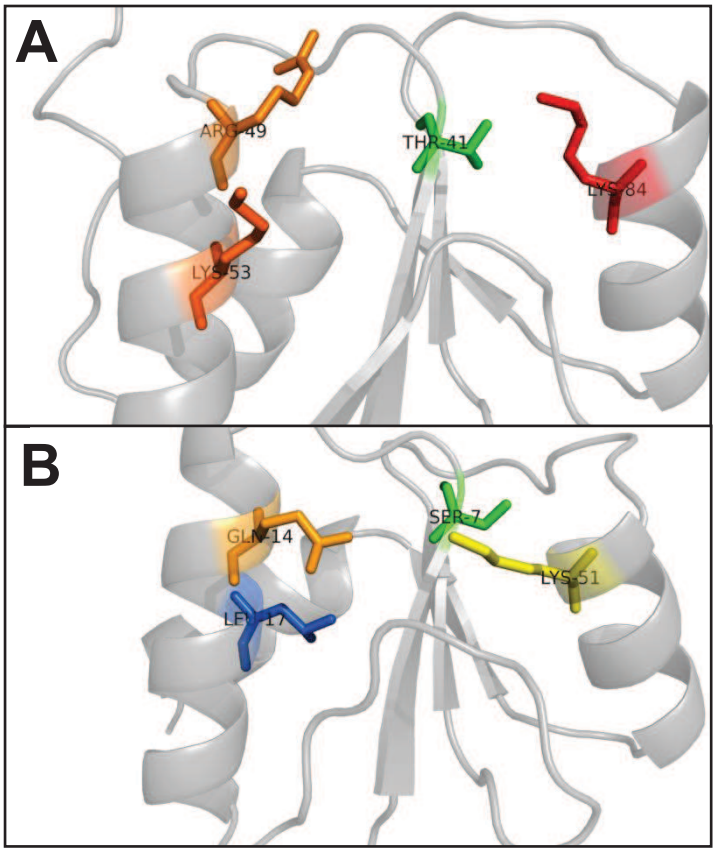


Abbildung 32: Vergleich der Energien der Bindungstasche einer N-terminalen d1 Domäne mit einer s1 Domäne. A: Energetisch eingefärbte Aminosäuren der Bindungstasche aus RFC1 (PDB-Id: 2K6G), welche in die s1 Gruppe eingeordnet ist. B: Energien der Bindungstasche von BARD1-N (PDB-Id: 2NTE).

8.4 Rückschlüsse der energetischen Analyse auf den evolutionären Verlauf der BRCT-Domäne

Zusammen mit den bisherigen aufgestellten Hypothesen und anhand der in dieser Arbeit dargelegten Ergebnisse über die energetischen und strukturellen Zusammenhänge zwischen den BRCT-Domänen, lässt sich für den gesamten evolutionären Verlauf dieser folgendes aussagen. Zum einen scheint es, dass sich die N-terminalen Domänen der d1 Gruppe direkt aus den Domänen der s1 Gruppe und somit noch vor den C-terminalen Domänen und denen der s2 und d2 Gruppe entwickelt haben. Dies belegt die große strukturelle und energetische Ähnlichkeit, sowie die geringen energetischen Anpassung, welche vollzogen wurden, um statt einen DNA Strang ein Peptid zu binden. Die C-terminalen Domänen entwickelten sich anschließend durch Verdoppelung der N-terminalen Domänen. Da in diesen jedoch keine Bindungsaktivität von Nöten war, sondern lediglich einige Reste für eine Stabilität des gebundenen Proteins sorgen sollten, fand innerhalb der Bindungstasche und der nicht mehr benötigten Bereiche (wie etwa die α2-Helix) große Veränderungen durch spontane Mutationen statt. Die Bildung der C-terminalen Domänen an sich war vermutlich deswegen nötig, da die Oberfläche eines gebundenen Proteins viel größer ist als jene eines DNA-Stranges und somit eine BRCT-Domäne für ein stabiles Binden nicht ausreichte. Zu der Zeit, in der in etwa die Verdoppelung der Domäne im Protein erfolgte, vollzogen sich vermutlich bei einigen noch eigenständigen N-terminalen Domänen weitere Veränderungen, welche eine neue Art der Proteinbindung hervorbrachten (die zukünftigen Domänen der s2 Gruppe). Diese benötigte die Aktivität der Bindungstasche nicht mehr, da sie das Ausbilden von Bindungen über die α1- und α3-Helix vorsieht. Da diese beiden Bereiche jedoch nach wie vor aktiv an der Bindung beteiligt waren, kam es zu ähnlichen Veränderungen wie bei den C-terminalen Domänen. Die Tatsache, dass in beiden Gruppen jedoch noch Reste der Aktivität der Bindungstasche vorliegen zeigt, dass sich deren Evolution in etwa zur gleichen Zeit vollzogen hat. Über die Entwicklung der d2 Domänen kann aufgrund von Datenmangel jedoch keine genaue Aussage gemacht werden. Dennoch ergeben sich anhand der Untersuchungen zwei Theorien über deren evolutionären Verlauf. Man kann annehmen, dass sich in diesen nach der Verdopplung der Domänen die Art der Bindung ausgebildet hat, welche bei den s2 Domänen zu beobachten ist. Dies kann dadurch belegt werden, dass in der untersuchten d2 Domäne (DNA Ligase IV) sowohl in der C- als auch der N-terminalen Domäne keine aktive Bindungstasche anzutreffen ist. Des Weiteren geht aus der Analyse von Studien hervor, dass diese Domäne einen sehr langen Linker aufweist und beide Domänen durch ausbilden von Bindungen durch ihre Helices an das entsprechende Gegenprotein binden. Die Art

der Bindung ist dabei denen der Domänen der s2 Gruppe sehr ähnlich. Theoretisch wäre es zudem möglich, dass durch eine ständige Verlängerung der Linkerregion zwischen der N- und C-terminalen Domäne es dazu gekommen ist, dass sich zwei voneinander unabhängige Domänen gebildet haben. So könnten sich einige Proteine entwickelt haben, in denen zwei BRCT-Domänen getrennt auftreten, wie es z.B. in XRCC1 der Fall ist. Um diese Theorien jedoch überprüfen zu können, müssten zusätzliche Analysen von d2 Domänen durchgeführt werden und in wie fern diese dann eine Verwandtschaft zu den restlichen Domänen aufweisen. Letztendlich zeigen die gesamten Untersuchungen dieser Arbeit, dass man den bisherigen evolutionären Verlauf der BRCT-Domäne weitestgehendste bestätigen kann und sich sogar neue mögliche Ansätze des evolutionären Verlaufes zeigen. Des Weiteren ist es dank der energetischen und strukturellen Analysemethode möglich, die verwandtschaftlichen Beziehungen zwischen den Domänen auf einer sehr genauen und detaillierten Art nachzuvollziehen. Jedoch war es leider nicht möglich energetisch markante Stellen zu identifizieren, welche z.B. zu einer genauen Aufspaltung der s2 und d1 Gruppen führte. Dennoch kann man anhand der jetzt erhaltenen Erkenntnisse eine unbekannte BRCT-Domäne genauer identifizieren in dem man sowohl die energetische Beschaffenheit der Bindungstasche betrachtet als auch die strukturellen Merkmale, welche für einige BRCT-Domänen typisch sind (z.B. die zusätzliche α-Helix bei C-terminalen Domänen).

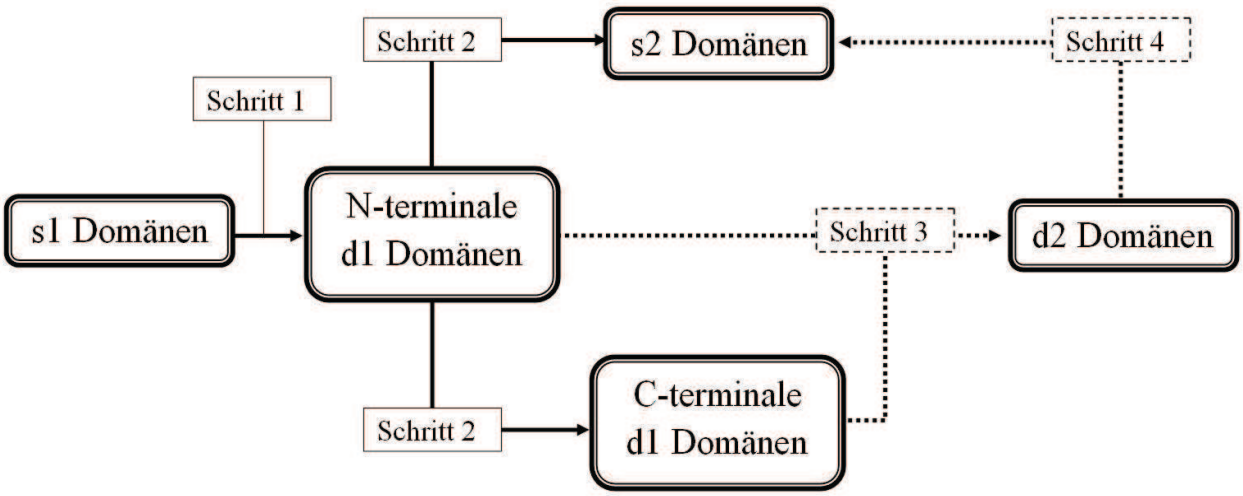


Abbildung 33: Schematische Darstellung des evolutionären Verlaufes, welcher sich anhand der erlangten Erkenntnisse ergibt. In Schritt eins erfolgte die energetische Anpassung der Bindungstasche. In Schritt zwei erfolgte zum einen die Verdoppelung der N-terminalen Domäne mit entsprechenden strukturellen und energetischen Anpassungen. Des weiteren erfolgten die gleichen Anpassungen ohne das eine Verdopplung der Domäne stattfand (Entstehung der s2 Domänen). Schritt drei und Schritt 4 sind die Hypothetischen Entwicklungen welche sich vollzogen haben könnten (siehe Text). Diese Schritte Bedarfen jedoch noch genaueren Untersuchungen.

9 Zusammenfassung

Die Untersuchungen dieser Arbeit haben gezeigt, dass es möglich ist mit Hilfe von Energieprofilen evolutionäre Verwandtschaftsbeziehungen zu rekonstruieren. Diese rekonstruierten Beziehungen stehen dabei in Korrelation mit bisherigen ermittelten Erkenntnissen, welche auf herkömmlicher sequenzieller Grundlage generiert wurden. Des Weiteren bieten die Energieprofile zusammen mit strukturellen Daten eine genauere und aufschlussreiche Grundlage für das Nachvollziehen der evolutionären Beziehungen. Das Anwenden von Energieprofilen auf phylogenetische Untersuchungen bietet somit nun zwei große Vorteile gegenüber dem Arbeiten auf sequenzieller Grundlage. Zum Einen lassen sich mittels Energieprofilen evolutionäre Verläufe einfacher rekonstruieren. Dies belegte die Tatsache, dass in den mittels distanzbasierenden Methoden generierten UPGMA-EP und NJ-EP Bäumen sich zum Teil verwandtschaftliche Beziehungen ablesen lassen, welche auf sequenzieller Grundlage mittels aufwendigeren charakterbasierenden Methoden generiert wurden. Eine derartige hohe Korrelation von Ergebnissen zwischen distanzbasierenden und charakterbasierenden phylogenetischen Methoden ist rein auf sequenzieller Grundlage nicht vorhanden. Zudem findet in den energetischen Bäumen eine bessere und logisch nachvollziehbarere Anordnung der Taxa anhand größerer markanter Eigenschaften statt (z.B. hohe strukturelle Ähnlichkeit oder energetische Ähnlichkeit innerhalb wichtiger funktioneller Bereiche). Der zweite große Vorteil ist, dass sich dank der kombinierten Analyse durch Energieprofile und Struktur die Veränderungen sowohl in funktionell wichtigen, als auch eher unwichtigen Bereichen sehr genau Nachvollziehen lassen. Somit können klarere und konkretere Aussagen darüber getroffen werden, inwieweit nun Veränderungen den evolutionären Verlauf beeinflusst haben und wie sich die Evolution genau vollzogen hat. Somit bieten Energieprofile eine bessere und genauere Grundlage für die Untersuchung evolutionärer Verwandtschaftsbeziehungen als Sequenzen.

Auf die BRCT-Domäne bezogen bedeutet dies, dass die Untersuchungen auf energetischer und struktureller Ebene zu genaueren Erkenntnissen über die Beschaffenheit der Funktion und den evolutionären Verlauf geführt haben. So konnte plausibel nachgewiesen werden, dass sich die N-terminalen Double Domänen mit ziemlicher Sicherheit aus den DNA bindenden Domänen der s1 Gruppen entwickelt haben. Ebenfalls wurde plausibel erläutert, dass sich die C-terminalen Domänen und die Domänen der s2 Gruppen parallel und in etwa zur gleichen Zeit entwickelt haben müssen. Zudem konnten neue Theorien

und Hypothesen aufgestellt werden, welche weitere evolutionäre Vorgänge genauer erklären würden, jedoch bedürfen diese noch einer genaueren Untersuchung und Beweisführung. So ist nunmehr, dank dem Einbeziehen energetischer Merkmale, ein grundlegenderes und klareres Bild der evolutionären Vorgänge der BRCT-Domäne vorhanden. Zudem konnten die funktionell wichtigen Bereiche und die bisher dazu erlangten Kenntnisse genau beschrieben werden, was wieder zu exakteren Erkenntnissen über die Funktion der einzelnen Domänen an sich führte. Letztendlich bietet diese Arbeit einen guten Grundansatz für weitere Untersuchungen der BRCT-Domäne für verschiedene Bereiche.

9.1 Ausblick

Anhand der Ergebnisse dieser Arbeit ergeben sich nun einige weitere Bereiche, welche es genauer zu erforschen und zu ergründen gilt. Zum Einen wäre es von großer Wichtigkeit, dass mehr Strukturen für zukünftige Analysen zu Verfügung stehen. Der momentane Mangel an zur Verfügung stehenden Strukturen macht den Einsatz von Energieprofilen auf größere Datenmengen bisher nicht möglich. Ein weiterer wichtiger Bereich wäre die Optimierung der phylogenetischen Methoden in Bezug auf das Arbeiten mittels Energieprofilen. Wie gezeigt wurde, ist es bisher nur möglich diese auf die distanzbasierenden UPGMA und NJ Methoden anzuwenden. Um nun eine Anwendung auf charakterbasierenden Methoden zu ermöglichen, müsste eine Art Substitutionsmatrix für Energieprofile entwickelt werden, welche mit bisherigen Substitutionsmatrizen wie PAM und BLOSUM vergleichbar ist. Sollte dies erfolgreich sein, so müsste anschließend untersucht werden, inwieweit sich die Ergebnisse zwischen diesen Methoden unterscheiden und welche Rückschlüsse daraus auf den evolutionären Verlauf und die Erkenntnisse dieser Arbeit gezogen werden können. Zudem wäre ebenfalls eine lokale Betrachtung des MEPAL möglich, was wiederum einen Vergleich zwischen lokal und global betrachteten UPGMA und NJ Bäumen mit sich führen würde. Des Weiteren wäre eine genauere Untersuchung der funktionell wichtigen Bereiche (wie in etwa der Bindungstasche) durch Simulierung der entsprechenden Bindungsvorgänge sehr aufschlussreich. Somit könnte überprüft werden, inwiefern sich einige energetische Veränderungen, welche beobachtet werden konnten (z.B. in der Bindungstasche oder am Ender der α1-Helix), auf die Funktion der Domäne auswirken. Daraus könnten wiederum Rückschlüsse gezogen werden, die den evolutionären Verlauf noch genauer beschreiben.

Literatur

[1] Biochemie, Eine Einführung für Mediziner und Naturwissenschaftler, Werner Müller-Esterl, Spektrum Akademiker Verlag, 1 Auflage 2004

[2] Heinke, Florian: *Energieprofilbasierende Analysemethoden von Proteinfamilien*, Mittweida, University of Applied Siences, Mathematik-Naturwissenschaften-Informatik, Bachelorarbeit, 2010

[3] Kaden, Janine: *Energieprofilbasierende Analyse der Pfam Familie 2-Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline Decarboxylase*, Mittweida, University of Applied Siences, Mathematik-Naturwissenschaften-Informatik, Bachelorarbeit, 2010

[4] Lexikon der Biochemie in zwei Bänden, Zweiter Band J bis Z, Übersetzung und Redaktion Angelika Fallert-Müller, Spektrum Akademischer Verlag GmbH Heidelberg, 2000

[5] Einführung in die Bioinformatik II, Entstehung von Strukturen, Vorlesungsskript Bioinformatik II, Dirk Labudde 05.03.2010

[6] Stöhr, Jan: *Biophysikalische Charakterisierung des Vorläufer- und Endzustandes von Fibrillen aus rekombinanten und natürlichen Prion-Proteinen*, Düsseldorf, Heinrich-Heine-Universität, Institut für Physikalische Biologie, Inaugural-Dissertation, 2007

[7] Bioinformatik, Eine Einführung, Arthur M. Lesk, Spekrtrum Aka-

demischer Verlag, Berlin, 2003

[8] Gene und Stammbäume, Ein Handbuch zur molekularen Phylogenetik; Volker Knoop & Kai Müller, Spektrum Akademischer Verlag Heidelberg, 2.Auflage, 2009

[9] Bioinformatische Algorithmen, Thema: Phylogenetische Bäume Teil 1-2 Vorlesungsskript Bioinformatik, Dirk Labudde, Mittweida, University of Applied Siences, Mathematik-Naturwissenschaften-Informatik, 19.03.2011

[10] Molekulare Phylogenie, Vorlesungsskript, Thomas Hankeln, Institüt für Molekulargenetik, Mainz, Gutenberg Universität, 2010

[11] Bioinformatik, Ein Leitfaden für Naturwissenschaftler; Andres Hansen, Birkhäuser Verlag, Basel-Boston-Berlin, 2004

[12] Understanding Bioinformatics; Marketa Zvelebil und Jeremy O. Baum, Garland Science, Taylor & Francis Group, LLC, 2008

[13] N. Saitou and M. Nei : *The neighbor-joining method: a new method for reconstructing phylogenetic trees*, Molecular Biology and Evolution. 1987, Vol 4(4), p406-425

[14] Joseph Felsenstein: *Evolutionary Trees from DNA Sequences: a Maximum Liklihood Approach,* Journal of Molecular Evolution, Springer Verlag 1981, 17:368-376

[15] URL <http://www.ncbi.nlm.nih.gov/pubmed/9034168> verfügbar

am 15.08.2012

[16] M. Kobayashi, Eiso AB, A. M. J. J. Bonvin, G. Siegal: *Structure of the DNA-bound BRCA1 C-terminal Region from Human Replication Factor C p140 and Model of the Protein-DNA Complex*, Journal of Biological Chemistry, 10087-10097, 26 März 2010, Vol. 285, No 13

[17] M. Kobayashi, F. Figaroa, N. Meeuwenoord. L. E. T. Jansen, G. Siegal: *Characterization of the DNA Binding and Structural Properties of the BRCT Region of Human Replication Factor p140 Subunit*, Journal of Biological Chemistry, 4308-4317, 17 Februar 2006, Vol. 281, No 7

[18] X. Zhang, S. More´ ra, P. A. Bates, P. C.Whitehead, A. I.Coffer, K. Hainbucher, R. A.Nash, M. J. E. Sternberg, T. Lindahl, P. S. Freemont: *Structure of an XRCC1 BRCT domain, a new Protein-Protein Interaction module*, 6404-6411, The EMBO Journal, 1998, Vol. 17, No. 21

[19] M. J. Cuneo, S. A. Gabel, J.M. Krahn, M. A. Ricker, R. E. London: *The structural basis for partitioning of the XRCC1-DNA ligase III BRCT-Mediated dimer complexes*, 7816-7827, Nucleic Acid Research, 2011, Vol. 39, No. 17

[20] Zi-Zhang Sheng et al: *Functional Evolution of BRCT Domains from Binding DNA to Protein*, Evolutionary Bioinformatics, 2011, 7:87–97

[21] Pei-Yu Wu et al: *Structural and Functional Interaction between the Human DNA repair Proteins DNA Ligase IV and XRCC4*, Molecu-

lar and Cellular Biology, June 2009, 3163-3172

[22] Charles Chung Yun Leung et al: *Molecular Basis of BACH1/FANCJ Recognition by TopBP1 in DNA Replication Checkpoint Control*, The Journal of Biological Chemistry, Vol. 286, No. 6, February 2011, 4292-4301

[23] Eugene F. et al: *Solution Structure of Polymerase μ's BRCT Domain Reveals an Element Essential for Its Role in Nonhomologous End Joining*, Biochemistry, October 2007, 46(43):12100-12110

[24] Mathieu Rappas et al: *Structure and function of the Rad9-binding region of the DNA-damage checkpoint adaptor TopBP*, Nucleic Acids Research, 2011, Vol. 39, No. 1, 313-324

[25] Charles Chung Yun Leung and J.N. Mark Glover: *BRCT domains Easy as one, two, three*, Cell Cycle, August 2011, 10:15, 2461-6470

[26] R. Scott Williams et al: *Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1*, Nature Publishing Group, 2001

[27] F. Heinke & D. Labudde: *Membrane Protein Stability Analyses byMeans of Protein Energy Profiles in Case of Nephrogenic Diabetes Insipidus,* Hindawi Publishing Corporation, Computational and Mathematical Methodes in Medicine, Volume 2012, Article ID 790281

[28] URL: <http://www.pdb.org/pdb/home/home.do> zuletzt verfügbar am 15.08.212

[29] Arnold K., Bordoli L., Kopp J., and Schwede T. (2006): *The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling*, Bioinformatics, 22,195-201

[30] Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009): *The SWISS-MODEL Repository and associated resources*, Nucleic Acids Research. 37, D387-D392

[31] Peitsch, M. C. (1995): *Protein modeling by E-mail Bio/Technology* 13: 658-660

[32] Florian Heinke: Java Programpaket, University of Applied Siences Mittweida, Dep. of Mathematics, Natural & Computer Sciences, 2012

[33] URL: <http://evolution.genetics.washington.edu/phylip.html>, Joe Felsenstein, Department of Genome Sciences and Department of Biology, University of Washington, Mail: joe@gs.washington.edu

[34] Andreas Prlic: *Biojava, an open source framework for bioinformatics in 2012*, Bioinformatics Advance Access, August 9, 2012 URL <http://biojava.org/wiki/Main_Page>, zuletzt verfügbar am 15.08.2012

[35] URL: <http://www.pymol.org/>, zuletzt verfügbar am 15.08.2012

Danksagung

An dieser Stelle möchte ich mich bei allen Bedanken, welche mir bei der Ausarbeitung dieser Arbeit zur Seite standen. An erster Stelle möchte ich Prof. Dr. Dirk Labudde dafür danken, dass er es mir ermöglichst hat an diesem Thema zu Arbeiten und dass er steht's für Fragen und Problemlösungen zur Verfügung stand. Das zweite große Dankeschön möchte ich meinem persönlichen Betreuer B.Sc. Florian Heinke zukommen lassen. Ohne die Unterstützung im Bereich der Programmierung, dem ständigen verfügbar sein bei offenen Fragen und Problemen und dem Anregen neuer Denkansätze wäre diese Arbeit in diesem Umfang vermutlich nicht zustande gekommen. Des Weiteren Bedanke ich mich bei meinen Freunden und Kommilitonen für anregende Diskusionen, welche nicht selten neue Einblicke und Ansichten lieferten, die Einfluss auf die Bearbeitung des Themas nahmen. Ein besonderer Dank gilt auch den Korrekturlesern, welche die Zeit und Mühen aufgebracht haben diese Arbeit nach Fehlern zu durchsuchen. Letztendlich möchte ich mich noch bei meinen Eltern und meinem Lebensgefährten für deren moralischen Beistand bedanken.

Anlagen

Teil 1 ..... A-I

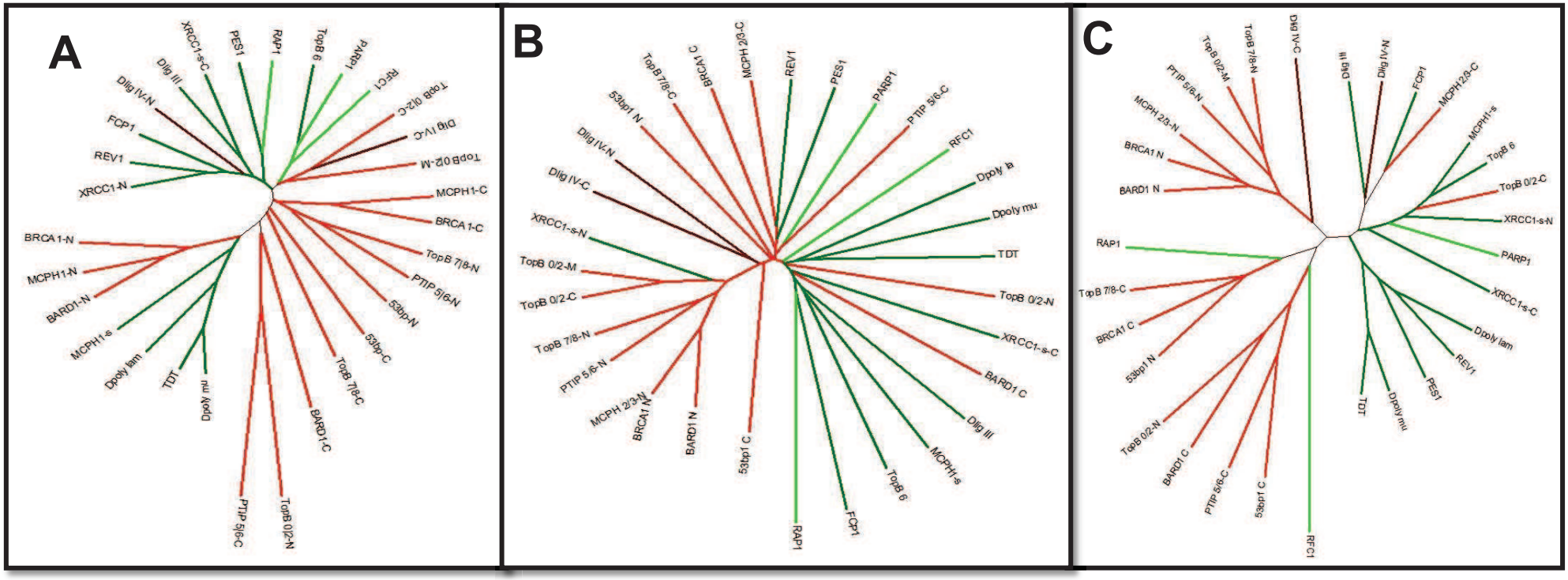
Teil 2 ..... A-II

Teil 3 ..... A-III

Teil 4 ..... A-IV

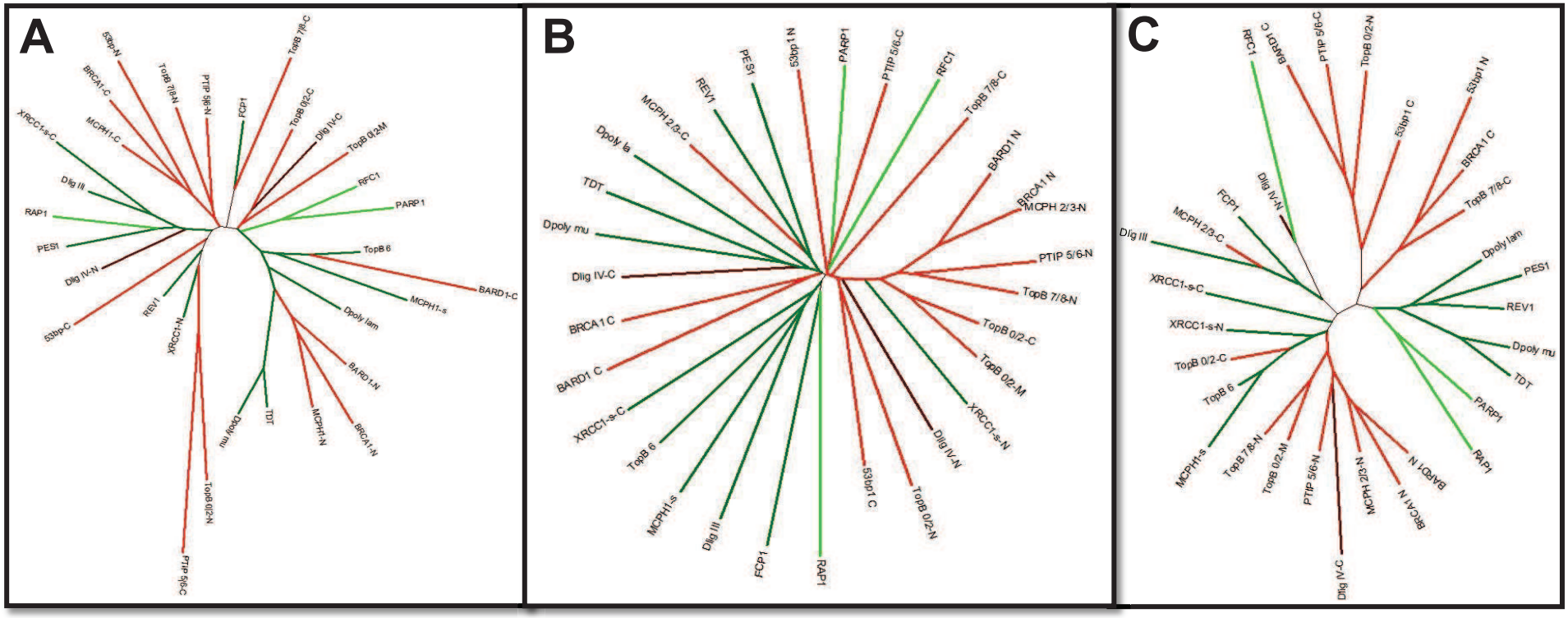
Anlagen, Teil 1

Die UPGMA Bäume der drei verschiedenen Grundlagen. A: UPGMA Sequenzbaum, B: UPGMA Strukturbaum, C: UPGMA Energieprofilbaum



Anlagen, Teil 2

Die Neighbor-Joining Bäume der drei verschiedenen Grundlagen. A: NJ Sequenzbaum, B: NJ Strukturbaum, C: NJ Energieprofil

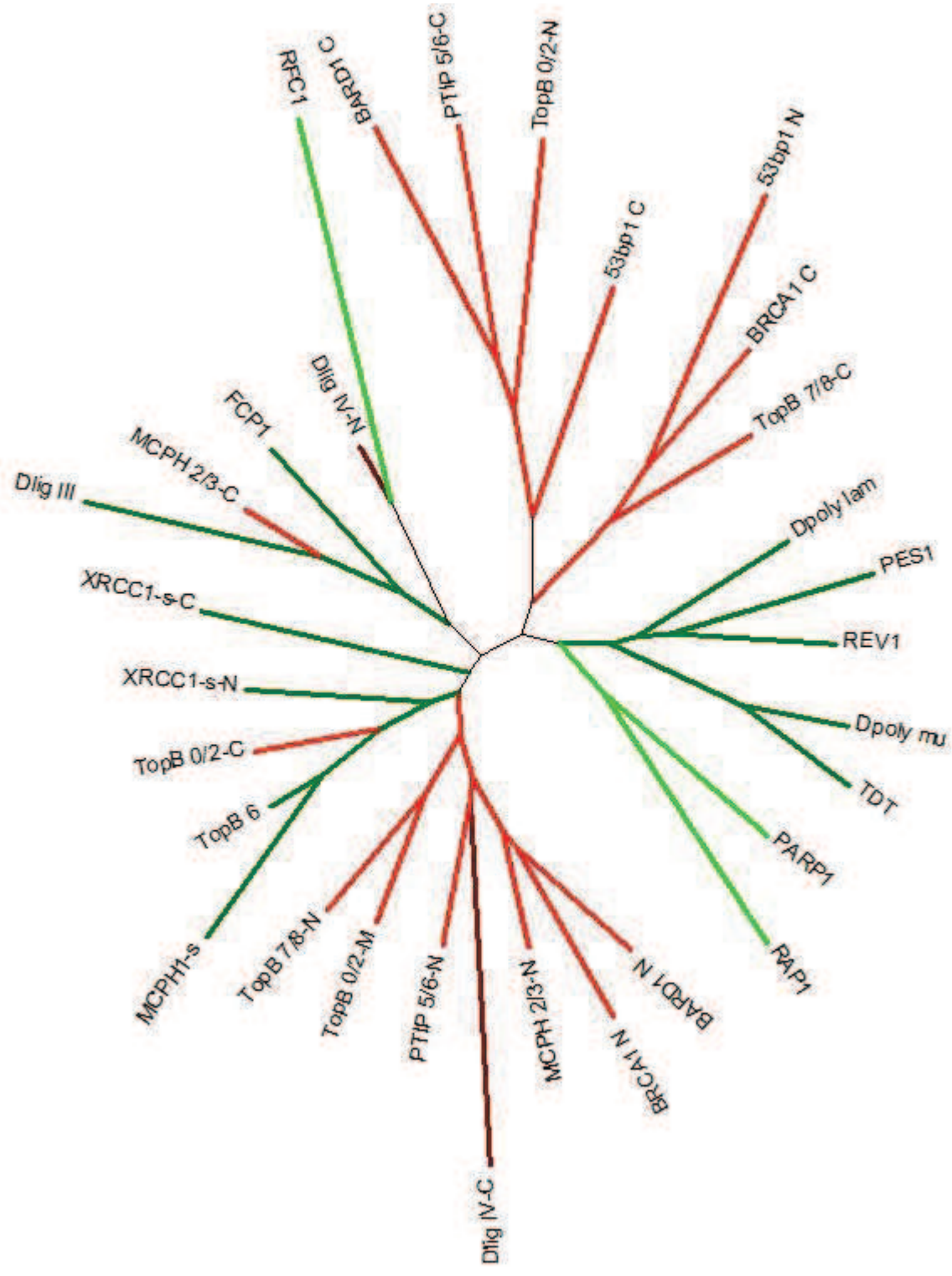


Anlagen, Teil 3

Die originalen Strukturbäume ohne logarithmierten pScore. A: NJ Baum, B: UPGMA Baum

Anlagen, Teil 4

Der auf Energieprofilen basierende Neighbor-Joining Baum (NJ-EP)



Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida den 16.08.2012

Mathias Langer